New computational protein design methods for de novo small molecule binding sites

by

James Edward Lucas

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
AND
UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:

DocuSigned by:
Tanja Kortemme
FA555ADA28F1439...

Tanja Kortemme

Chair

DocuSigned by:
John Dueber
DocuSigned by:
B...

John Dueber

DocuSigned by:
Michael J Keiser

Michael J Keiser

DocuSigned by:
Andrej Sali
8F7A6AB94F2C4F4...

Andrej Sali

Committee Members

Acknowledgements

Upon reflection, I am incredibly grateful to have had the opportunity to complete a PhD in such an exciting field of study. This would not have been possible without the support of my friends, family, and mentors, each of whom contributed to the success of my PhD in their own way.

I would like to thank my advisor, Tanja Kortemme, for taking me into her lab and allowing me to pursue ambitious research projects. I also want to thank the members of the Kortemme lab for sharing their expertise with me. Special thanks to Kale, Kyle, Xingjie, and Shane for answering my incessant questions; my dissertation ended up being purely computational and my research was only possible because of your help. I'm delighted that Kyle and I are still able to catch up over lunch every month and that Kale and I still talk science even now that he has moved across the country. I also want to thank Anum for providing me with much needed perspective and support during our ice cream and boba walks.

A special thanks to John for investing so much of his time into me. I am hugely grateful for the time you set aside to chat (whether or not it was scheduled) and your support for potential research projects. It was a pleasure developing the protein engineering course with you and I hope that we can continue to work together on the plasmid database! I also want to thank Stacy for being a wonderful industry mentor over the past year. It has been a delight touching base every month and learning about exciting research outside of academia. My appreciation for your encouragement to pursue new opportunities and going out of your way to introduce me to your industry contacts cannot be overstated.

I was only ever in the position to pursue a PhD because of my outstanding undergraduate mentors who continued to support me throughout graduate school. Thanks to Justin and Marc who continued to make time for me when I visited Davis and provided me with research and career advice throughout my PhD. Justin is the reason I got into protein design in the first place, and his mentorship continues to drive me to pursue new and challenging research endeavors. I also want to thank members of the Siegel lab with whom I have continued to have fruitful conversations about science and life in general: Wilson, Steve, Terry, YouTian, and Katie. It has been a pleasure keeping in touch and watching each of you advance through your careers and life goals.

I want to thank my parents for always wanting the best for me. Even during my PhD, they have

continued to help me emotionally and financially. At the end of the day, I could always count on you to look out for my well being. I am also deeply indebted to my close circle of friends who continued to support me throughout my journey. It is impossible to put into words how each of you helped propel me forward, but I'm confident that each and every one of you knows how much our friendship means to me.

Kendra, it pains me knowing that I cannot share this moment with you. My heart aches every time I realize that we will never be able to follow through on our life plans together. I think about you every day and it still hurts every time. I constantly think back to our last conversation on your birthday and how excited you were to start applying to graduate school. I wish I had your enthusiasm for science and learning, and I do my best to embrace your zeal for life every day. This work is dedicated to you.

New Computational Protein Design Methods for De Novo Small Molecule Binding Sites

James Edward Lucas

Abstract

Protein binding to small molecules is fundamental to many biological processes, yet it remains challenging to predictively design this functionality de novo. Current state-of-the-art computational design methods typically rely on existing small molecule binding sites or protein scaffolds with existing shape complementarity for a target ligand. Here we introduce new methods that utilize pools of discrete contacts observed in the Protein Data Bank between protein residues and defined small molecule ligand substructures (ligand fragments). We use the Rosetta Molecular Modeling Suite to recombine protein residues in these contact pools to generate hundreds of thousands of energetically favorable binding sites for a target ligand. These composite binding sites are built into existing scaffold proteins matching the intended binding site geometry with high accuracy. In addition, we apply pools of rotamers interacting with the target ligand to augment Rosetta's conventional design machinery and improve key metrics known to be predictive of design success. We demonstrate that our method reliably builds diverse binding sites into different scaffold proteins for a variety of target molecules. Our generalizable de novo ligand binding site design method will lay the foundation for versatile design of protein to interface previously unattainable molecules for applications in medical diagnostics and synthetic biology.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Proteins are macromolecules that perform the majority of life-sustaining functions within cells. These functions are extremely diverse and include performing chemistry, sensing and responding to stimuli, transporting cargo, regulating transcription and translation of genetic material, and providing structural support. Many of these processes rely on molecular recognition and binding of small molecules. However, despite comprehensive understanding of the physicochemical properties that proteins use to mediate selective high-affinity interactions with small molecules, current methods still lack the ability to predictively design cavities that impart binding function for specific ligands. The ability to design proteins that bind to small molecules of interest is essential to designing enzymes and sensors for a variety of applications. The work in this dissertation in particular was originally motivated by the design of chemically-induced protein dimerization systems to couple sensing of arbitrary small molecule targets to a wide array of in vitro and in vivo responses.

To this end, there has been extensive efforts to develop computational methods for protein small molecule binding site design. The Rosetta Macromolecular Modeling Suite is one such computational tool that has been successfully applied to model and impart a variety of protein structures and functions, including ligand binding. While there have been several successful methods developed within Rosetta for designing small molecule binding sites, there are still several deficiencies that need to be addressed before predictable design of small molecule binding sites can become routine.

The work in this dissertation attempts to address several of these deficiencies by leveraging

empirical protein-ligand contact information from the Protein Data Bank to augment the design capabilities of Rosetta. This work aims to develop a framework where binding site design can be routinely performed for arbitrary small molecule targets.

Chapter 2

Computational Design Methods for De Novo Small Molecule Binding Sites

2.1 Introduction

Despite significant advances in de novo design of protein structures, innovations in algorithms and methodologies for the computational design of protein function have not kept pace.[1,2] In particular, it remains challenging to design proteins that bind new small molecule ligands.[3,4] The ability to design proteins with high-affinity interactions for defined small molecule targets is important for creating enzymes with novel substrates, receptors and transcription factors that sense and respond to unique inducer molecules, and binders that can recognize and sequester arbitrary ligands. On-demand design of proteins with defined small-molecule binding functionality would have many applications in bioremediation, synthetic biology, and medical diagnostics.

Various strategies have been developed and successfully applied to the computational design of ligand binding function. For instance, design methods in the Rosetta Macromolecular Modeling Suite[5] have been used to reengineer proteins to bind digoxigenin,[6] fentanyl,[7] and 17a-hydroxylprogesterone,[3] to create a new binding site for the metabolic intermediate farnesyl pyrophosphate in a protein-protein interface to build synthetic sense/response systems,[8] and to place a binding site for (Z)-4-(3,5- difluoro-4-hydroxybenzylidene)-1,2-dimethyl-1H-imidazol-5(4H)-one (DFHBI) into the cavity of a de novo designed beta-barrel.[9] There are several examples of de

novo designed helical bundles that bind various ligands,[10] including a recent synthetic porphyrin binder that was engineered through simultaneous optimization of the binding site and a remote, well-packed core.[11] Another common avenue for designing binders with desired function is by changing the ligand specificity of an existing binder, for example by grafting a binding site of a close homolog.[12]

Common design protocols to introduce a new ligand binding function into a protein[13] begin with the specification of the binding site, where amino acid side chains and conformations are geometrically defined to make key interactions with the target ligand. These binding site definitions are typically derived based on chemical intuition or taken from an existing protein in complex with the target ligand present in the Protein Data Bank (PDB). In a second step, the target ligand and its defined contacts are placed into a new protein (termed "scaffold") using geometric matching, such as the RosettaMatch application.[14] Given a set of geometric constraints defining side chain interactions with a target ligand, RosettaMatch attempts to find backbone positions in a scaffold protein where side chains can be placed to satisfy all protein-ligand contact constraints and places the ligand within the protein subject to these constraints. In a third step, sequence positions surrounding any successfully matched binding site are redesigned to accommodate the newly introduced ligand and defined side chains. The resulting design models are ranked based on predicted protein stability and interaction energy with the ligand, in addition to a range of design filters, and predictions are selected to be experimentally validated.

Despite previously demonstrated success of these protocols,[6,8] there are still several limitations that prevent generalizable de novo design of ligand binding sites. First, binding site geometries need to be predefined for a target ligand. Even when binding site definitions can be derived from existing complexes for the target ligand in the PDB, there may only be a few examples (if any) and these might be limited to low-affinity binding interactions. Second, geometric matching algorithms are typically only capable of incorporating binding site definitions composed of 3-5 residues due to the complexity of finding backbone geometries in scaffold proteins that satisfy all user-defined constraints. To find solutions, it is necessary to relax these constraints at the expense of deviating significantly from the geometries in the binding site definition. Third, binding site definitions matched into a scaffold protein only constitute a part of the binding site and the remainder of the binding site environment needs to be optimized through algorithms that change sequence identity

(design) or side chain conformations (rotamer packing).[15,16] Common binding site design methods such as those implemented in Rosetta, including the recently developed Rotamer Interaction Field,[9] rely on Rosetta's energy function to introduce favorable interactions with the ligand. However, limitations in the energy function used for design may fail to capture important interactions with the target ligand, such as pi-cation and pi-pi interactions[17] or interactions with atom types that are commonly encountered in ligands but less well parameterized.[18] Finally, filtering steps after design typically eliminate a majority of design candidates because they possess poor shape complementarity, ligand burial, hydrogen bond satisfaction, and other metrics that are predictive of success.

Here we describe a new approach that uses protein-ligand contact observations in the PDB to address current shortcomings in defining and designing ligand binding sites. Our approach generates hundreds of thousands of new binding site definitions in an automated fashion for arbitrary target ligand conformations, regardless of whether a complete binding site definition can be derived from an existing complex in the PDB. These binding site definitions yield thousands of RosettaMatch solutions that agree well with originally defined geometries. We also introduce a new design method that recapitulates key interactions in ligand binding sites. Moreover, we show that the new design method improves several metrics in design that are predictive of success when designs are experimentally characterized. While we have incorporated these new methods in Rosetta, the principles are generally applicable to other design approaches. The methods introduced here will have broad utility to design binders for arbitrary small molecules, towards predictable design of small molecule sensors, inducible transcription factors, and other functional proteins.

2.2 Results

We sought to leverage the wealth of information in the PDB to address current shortcomings in existing protein-ligand interface design methods. While high-affinity protein-ligand complexes for small molecules of interest may be sparse or non-existent, our approach is built on the following: (i) Target ligands are decomposed into substructures for which there are many contacts in the PDB (Figure 2.1 panel A), and (ii) these observed contacts with each ligand-derived substructure are

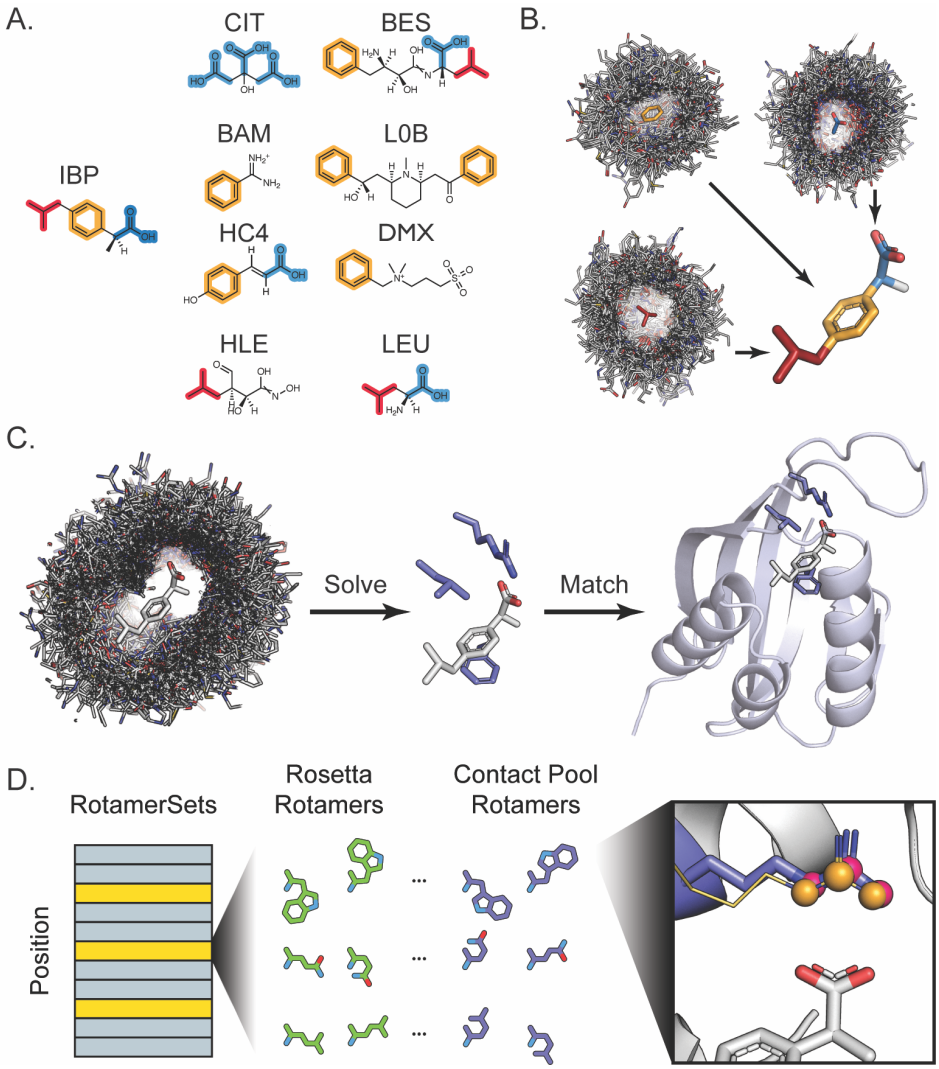


Figure 2.1: Protocol Overview. A. Target small molecules are decomposed into fragments (highlighted red, orange, and blue) that are present as a substructure in a wide range of molecules bound to proteins in the PDB. B. Protein complexes bound to substructure-containing molecules are systematically parsed to generate contact pools representing all contact modes with each fragment present in the PDB. These contacts are mapped onto the full target ligand (here ibuprofen) to create a conformer-specific contact pool for downstream steps. C. A simulated annealing protocol is used to assemble hundreds of thousands of three-residue composite binding sites from a target ligand contact pool. RosettaMatch is then used to find scaffold proteins that can geometrically accommodate a composite binding site solution. D. Sequence design is performed by Rosetta's design machinery, where positions in the protein can be set to designable (yellow bars) or not designable (grey bars). Contact pools are used to supplement the set of rotamers generated by Rosetta for designable positions. Rotamers are built in the context of a potential ligand interface and any rotamers (purple sticks, pink spheres) that recapitulate an interaction in the contact pool (gold lines, orange spheres) are added to the standard RotamerSet (green) and provided a score bonus with the special_rot score term.

used to assemble a pool of contacts for the target ligand (Figure 2.1 panel B).

We reasoned that these contact pools should improve design of new ligand binding sites in three principal ways. First, reassembling "composite" binding sites by recombining protein-ligand substructure contacts should provide an automated method to generate a large number of potential binding site definitions (even for ligands for which no complete binding site definition exists in the PDB) that can be matched into scaffold proteins. Second, by dramatically increasing the number of binding site definitions used in the geometric matching step, we should increase the number of scaffold protein "hits" that can accommodate these geometries well. Third, contact pools should also be useful in the design step, by incorporating protein-ligand contacts that might otherwise be missed in conventional design. As a result, we reasoned that design with contact pools should increase key metrics such as protein-ligand shape complementarity in the design filtering step. To test these ideas, we first apply the contact pools of protein-ligand interactions to assemble millions of new binding site definitions that can be incorporated into protein scaffolds with RosettaMatch (Figure 2.1 panel C). Second, we use contact pools to augment and inform existing design protocols (Figure 2.1 panel D).[19] In the following, we first describe the implementation of our protocol to assemble contact pools. We then show that the application of contact pools to inform design yields millions of strictly defined binding sites and improves key design metrics

2.2.1 Generation of contact pools

Our protocol produces a set of contact pools for an arbitrary number of potential ligand conformers that can be applied to either A) generate binding site definitions that can be used as constraints for the RosettaMatch application or B) generate rotamers to augment the conventional Rosetta design machinery (i.e. the Packer).

We define a "fragment" as a ligand substructure constituting a distinct chemical moiety (Figure 2.1 panel A) that will form an interaction with a protein (Figure 2.1 panel B). Fragments are composed of at least three atoms and are rigid substructures of the target ligand. While several automated small molecule fragmentation methods exist, these methods typically break bonds for retrosynthetic analysis in the context of fragment-based drug discovery.[20,21] The assembly of contact pools instead relies on the chemical intuition of the user to identify fragments that will mediate

important interactions at the protein-ligand interface (e.g. hydrogen bond donors/acceptors, ring systems) but are not necessarily segmented by breakable bonds.

2.2.2 The PDB is rich in protein-fragment contact information

To demonstrate the number and diversity of unique types of contacts that may be observed with fragments in the PDB, we generated 34 fragments for a variety of chemical moieties that are found in common drugs, toxins, and metabolites (Figure 2.2 panel A, Table 2.3). All protein-ligand complexes in the PDB that contain these fragments were retrieved and transformed to superpose the ligand substructure onto a reference fragment. Only protein side chain contacts within 4Å of the ligand were kept. We defined a 16-dimensional feature vector for each protein residue encoding the type of chemical contact mediated between the residue and the fragment (see Methods). We then applied hierarchical agglomerative clustering to generate clusters of protein contacts within the contact pool that we define to mediate unique "contact modes" (i.e. clusters) with each ligand fragment.

Sorting clusters by occupancy revealed that fragments often possess a small number ($<$5% of all clusters) of preferred, high-occupancy contact modes with remaining clusters consisting of less prevalent contact modes (Figure 2.2 panel B). This observation suggests that proteins in the PDB exhibit preferred modes of mediating interactions with the chemical environments that define each fragment. Indeed, visually inspecting clusters with the highest occupancies for various fragments revealed residue types mediating well-defined, favorable contact geometries that one would expect. For instance, the most common contact modes with the adenine fragment (second row from top) consist of bidentate hydrogen bonding interactions with the base mediated by side chain functional groups (e.g. asparagine/glutamine carboxyamide) as well as backbone carbonyl/amine functional groups (Figure 2.2 panel D).

The preference for a select number of contact modes led us to investigate what proportion of the PDB would need to be sampled to recover a majority ($>$80%) of contact modes for each fragment. To address this question, we bootstrapped 1000 random samples for different proportions of the PDB and counted the number of unique contact modes recovered as defined by our previously generated clusters. Only a minority fraction of the PDB (typically $<$40%) was required to recover

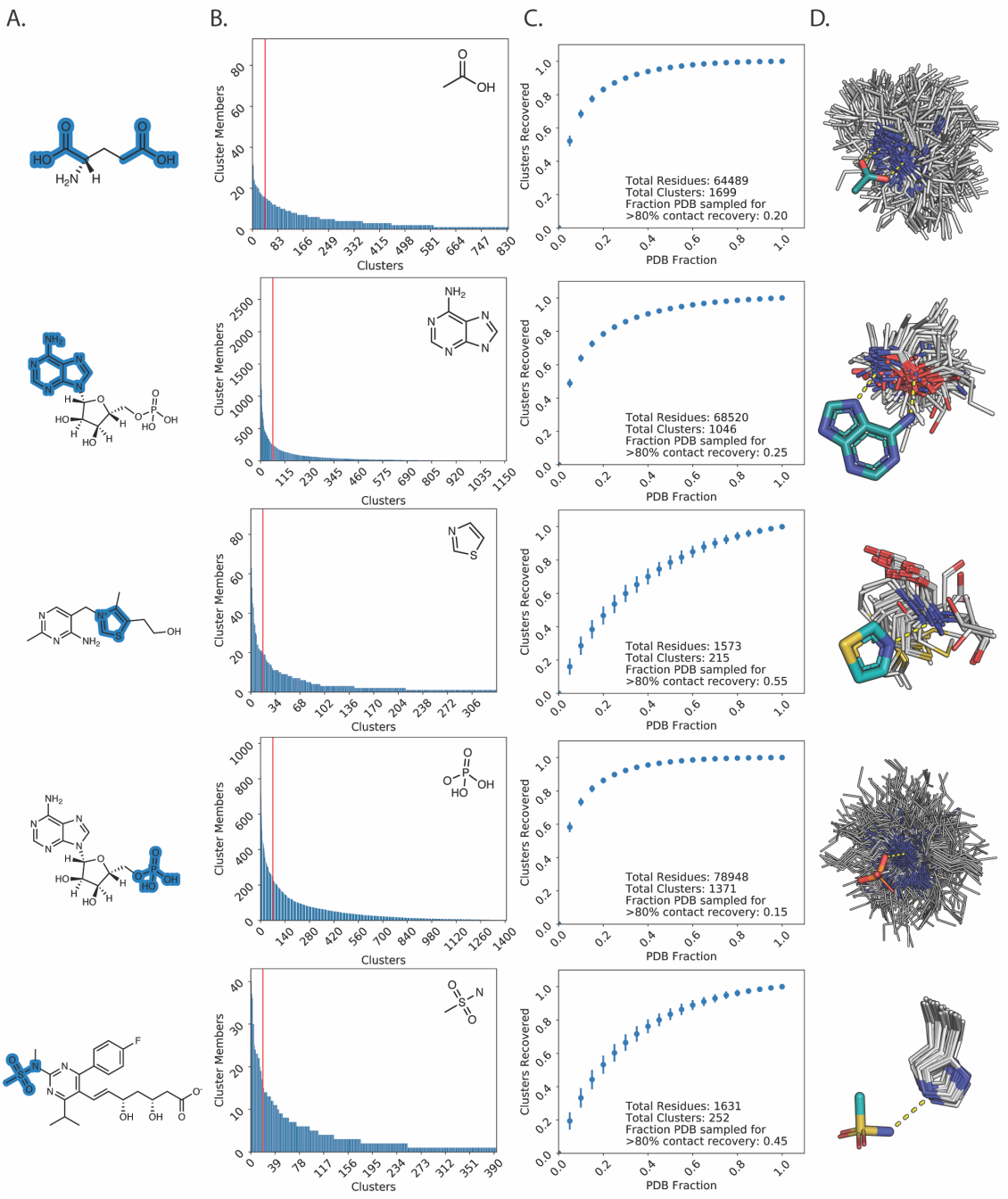


Figure 2.2: PDB sufficiently samples ligand-fragment contact space. A. Examples of fragment definitions (highlighted in blue) and context as a substructure of a small molecule. B. Clusters for each fragment are plotted in order of the number of members for each cluster ("contact mode"). Clusters to the left of the red vertical line represent the top 5% (high-occupancy) contact modes. C. Average and standard deviation for the number of contact modes recovered at given fractions of the PDB. D. Examples of high-occupancy clusters for fragments that illustrate expected favorable contact modes. White lines: clustered side chain contacts, teal: fragment carbons, orange: phosphorous, yellow: sulfur, red: oxygen, blue: nitrogen. Hydrogen bonding contacts are depicted as yellow dashed lines.

at least 80% of unique contact modes with each fragment (Figure 2.2 panel C).

2.2.3 Using composite binding sites increases the number of high-quality matches

Our analysis of contact modes for ligand fragments in the PDB demonstrates the diversity of side chain identities and geometries that proteins use to mediate interactions with unique chemical moieties on the ligand. We next sought to use this diversity to expand the number of side chain compositions and geometries that could be used to define a ligand binding site for methods like RosettaMatch. Instead of using an existing binding site extracted from the PDB to define side chain interactions with a target ligand, we used protein-fragment contacts to create "composite" binding sites. We first fragmented the target ligand and generated a contact pool for each fragment using discrete side chain-fragment interactions in the PDB. We then transformed the orientation of all protein residues in each fragment contact pool relative to the source fragment in the target ligand to preserve observed contact geometries (Figure 2.1 panel B). For the assembly of composite binding sites, we filtered these contacts using the fa_rep (Lennard-Jones, repulsive component), fa_atr (Lennard-Jones, attractive component), fa_elec (coulombic electrostatic potential), hbond_sc (side chain-side chain hydrogen bonds[22]), hbond_bb_sc (backbone-side chain hydrogen bonds), and fa_sol (Lazaridis-Karplus solvation[23]) two-body terms in the Rosetta REF2015 all-atom energy function.[17] This additional filtering step yields a collection of residue contacts with the target ligand that are not only observed in the PDB to interact with the target ligand, but are also determined to be energetically favorable by Rosetta, to serve as "hot spot" residues for composite binding sites.[24] These filtered pools contained on average 2,800 unique contacts for each ligand. This procedure can be repeated for different ligand conformers generated using a conformer search tool such as OpenEye Omega[25] or RDKit.[26]

Given a contact pool for a ligand conformer, we applied a simulated annealing Monte Carlo protocol similar to the Rosetta side chain rotamer optimization method (Packer) to yield low-energy three-residue composite binding site solutions for the target ligand (Figure 2.1 panel D) (here we illustrate the method with three-residue binding sites but larger sites could be used). Up to ten trajectories are attempted for each contact pool and the best 100,000 three-residue binding sites for each trajectory are recorded. The results from separate trajectories are consolidated and the

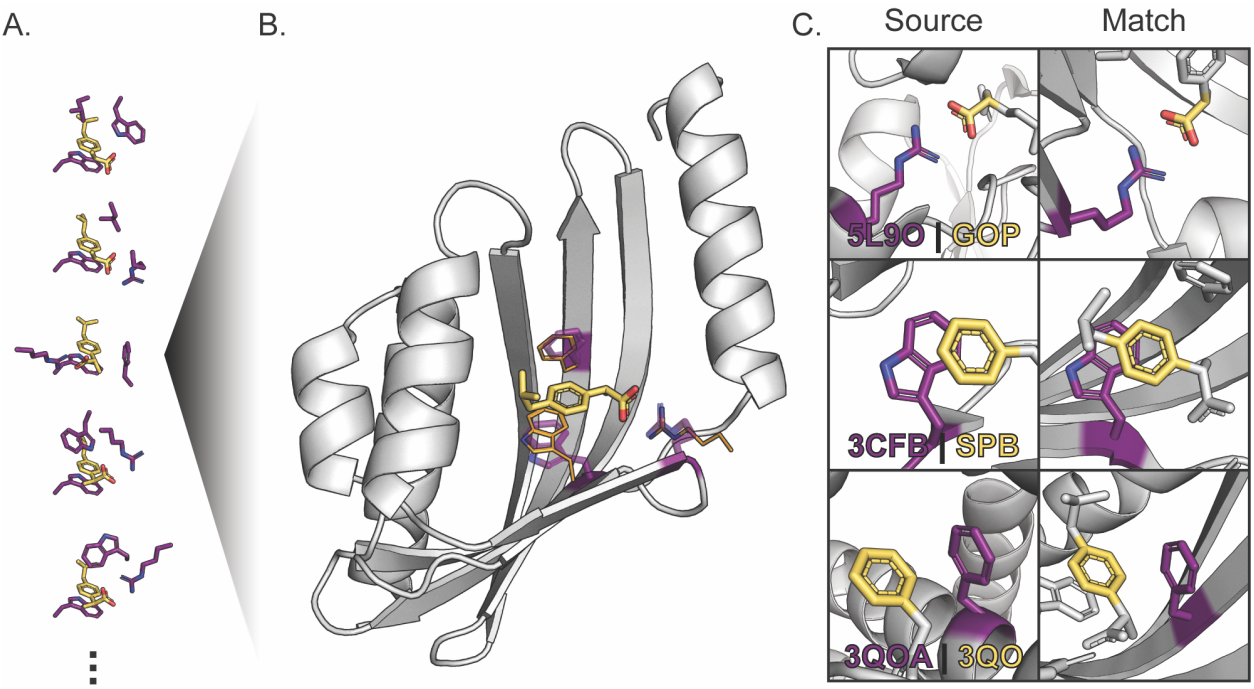


Figure 2.3: Discrete Protein-Fragment Contacts are Assembled into Composite Binding Sites. A. Thousands of composite binding sites are generated by combining discrete, observed contacts with fragments that compose the target ligand into low-energy configurations using a simulated annealing protocol. B. RosettaMatch finds scaffold proteins with existing cavities that can accommodate the target ligand with close adherence to contact residue geometries defined in composite binding site solutions (RosettaMatch residues: purple sticks, composite binding site definition: orange lines). C. Comparison of protein residue (purple) interactions with defined fragments (yellow) in the context of a protein-ligand complex that served as the source of a contact pool interaction (Source) and in the context of a match produced using RosettaMatch and geometric constraints derived from a composite binding site solution (Match). The source protein-ligand complex PDB ID and ligand chemical component identifier are provided for each source interaction.

Table 2.1: Numbers of matches found comparing complete binding sites extracted from the PDB and composite binding sites generated by our method. Filtered match counts constitute the 25 best-scoring matches that pass quality filters for each set of matches that result from constraints for a single composite binding site and place the same residue identities at the same positions in a scaffold protein.

	Complete Binding Sites From PDB			Solved Constraints From Contact Pools		
Chemical Component Identifier	# Complete Binding Sites	Raw Matches	Filtered Matches	# Constraints	Raw Matches	Filtered Matches
38E	2	2141	55	5000	2430929	31430
AFN	0	0	0	5000	1560062	27806
ATZ	1	0	0	5000	3797641	22525
DOG	4	3175	94	5000	578060	5041
IBP	12	1267	83	5000	1550018	16206
IM4	2	0	0	5000	3605143	49679
LFN	2	329	53	5000	4575191	38794
NPS	7	122	20	5000	2473723	18229

5,000 lowest-energy unique composite binding site solutions are used to generate constraints for RosettaMatch.

Next, we used RosettaMatch to build composite binding sites into a monomer scaffold set consisting of 401 proteins that had been previously applied to successfully redesign a protein to bind a new ligand.[6] We use stringent RosettaMatch constraints (see Methods) to ensure that the match solutions found by RosettaMatch do not significantly deviate from defined constraint geometries (Figure 2.3). We generated 5,000 composite binding sites for eight ligands (Table 2.4) and compared the number of matches found to those found using constraints generated from existing binding sites for the corresponding ligands in the PDB. In every case, using composite binding site solutions produces $>$50-fold more matches than conventional constraints derived from existing protein-ligand complexes (Table 2.1). Only 5000 of the lowest-scoring composite binding sites were used for this benchmark; many more potential match solutions can be found with the hundreds of thousands of composite binding site solutions generated with this method. Moreover, we find many more matches with composite binding sites that pass binding site energy, bump check, and designability quality filters (see Methods, "filtered matches" in Table 1) than when using PDB-derived constraints. Increasing the number of high-quality matches in this step is important since they serve as starting points for design, and, frequently, even good matches do not yield high-quality final design models because of difficulties generating good binding site environments in the design step. We note that our method can generate composite binding sites and large numbers of matches for ligands for which protein-ligand complexes are rare or do not exist in the PDB (e.g. chemical component identifiers ATZ, AFN in Table 1).

2.2.4 Complementary Rotamers Improve Binding Site Recapitulation

In addition to improving the numbers and quality of generated matches using composite binding sites, we hypothesized that contact pools could also be used to improve the design step. Since contact pools possess a wealth of information on how proteins in the PDB interact with different chemical moieties present on ligands, we sought to incorporate this information into Rosetta's core design machinery (the Packer). We used the backbone-dependent Dunbrack library[16] to generate rotamers for protein positions at the ligand interface and added complementary rotamers that recapitulate interactions in the contact pool to the Packer RotamerSets (Figure 2.1 panel C). These complementary rotamers were generated with additional χ-angle sampling for χ_1 through χ_4. For each rotamer built at a position within a binding site, a set of three contact atoms were defined based on contact geometry with the ligand. These atoms were compared to the same atoms for the same residue identities in the contact pool. If rotamer contact atoms achieved an RMSD of $\leq$1.5Å with any residue in the contact pool, it was added to the set of complementary rotamers for the current position. These additional rotamers are flagged as a "special_rot" variant.[19] An additional "special_rot" score term is enabled with a customizable bonus to bias incorporation of empirically determined residue-ligand contacts during design.

To determine an appropriate value for the special_rot score bonus, we used a native sequence recovery test[15,27,28,29] on a panel of protein-ligand complexes. We identified a total of 22 protein-ligand complexes from the BindingMOAD database[30] and generated contact pools for each unique ligand (Table 2.5) to create a set of empirically-determined rotamers to complement the rotamers generated by the Packer. These "complementary rotamers" were included in 5000 design trajectories for special_rot bonus values ranging from 0 (complementary rotamers are added to the Packer but no score bonus is applied) to -10 Rosetta Energy Units (REUs). For each complex, first and second shell contacts with the ligand as determined by the Rosetta Neighborhood and ClashBasedRepackShell selectors (see Methods) were subjected to design. Profile similarity (see Methods) and sequence recovery to the native complex sequence were calculated for designable positions in the benchmark complexes to determine whether complementary rotamers improve Rosetta's ability to design the binding site environment, and if so, which special_rot score bonus value was optimal. For the purposes of the native sequence recovery benchmark, a design posi-

tion is considered recovered if >50% of designs incorporate the residue identity observed in the original protein-ligand complex.

The application of complementary rotamers with the special_rot score term improves, or at least matches, native sequence recovery over the unmodified Packer for special_rot bonus values up to -4.0 REU in the benchmark set (Figure 2.4 panel A), with a special_rot bonus of -1.5 REU providing optimal sequence recovery in this test. When considering the median profile similarity for all 589 designable positions across the 22 complexes in our benchmark set as a metric, it initially appears that the application of complementary rotamers do not provide a significant improvement to design (Figure 2.8). Median profile similarity remains constant with a special_rot bonus between -0.5 and -4.0 REU, and is comparable to profiles generated using Rosetta's unmodified Packer as well as the addition of complementary rotamers without the application of the special_rot score term. However, a significant fraction of designable positions achieved high profile similarity (>0.9) regardless of modifications to the Packer. These positions were removed from subsequent analysis to investigate the impact of complementary rotamers. With these positions removed, it becomes evident that the addition of complementary rotamers can indeed provide a modest improvement in median profile similarity above a special_rot bonus of -5.0 REU (Figure 2.4 panel B). As with sequence recovery, a special_rot bonus of -1.5 REU appears optimal for median profile similarity in this test. Beyond a special_rot bonus of -4.0 REU, the addition of the complementary rotamers becomes detrimental for both the native sequence recovery and profile similarity metrics (Figure 2.4 panel A,B). This behavior is expected as the special_rot bonus begins to outweigh penalties otherwise incurred due to unfavorable physical interactions (e.g. steric clashes penalized by the repulsive score term fa_rep).

To investigate the per-position contributions of the complementary rotamers, we compared position-specific profile similarities for all designable positions in our benchmark set for three special_rot bonus values to results using the unmodified Packer (Figure 2.4 panel C). The addition of complementary rotamers alone (with a special_rot bonus = 0) does not have a notable impact on profile similarity, although two positions achieve >0.9 profile similarities with the complementary rotamers while the unmodified Packer only achieves profile similarities <0.5 for the same positions (Figure 2.4 panel C, left). The improvements provided by the complementary rotamers become apparent with a special_rot bonus of -1.5 REU, where 56 of 589 design positions achieve >0.1

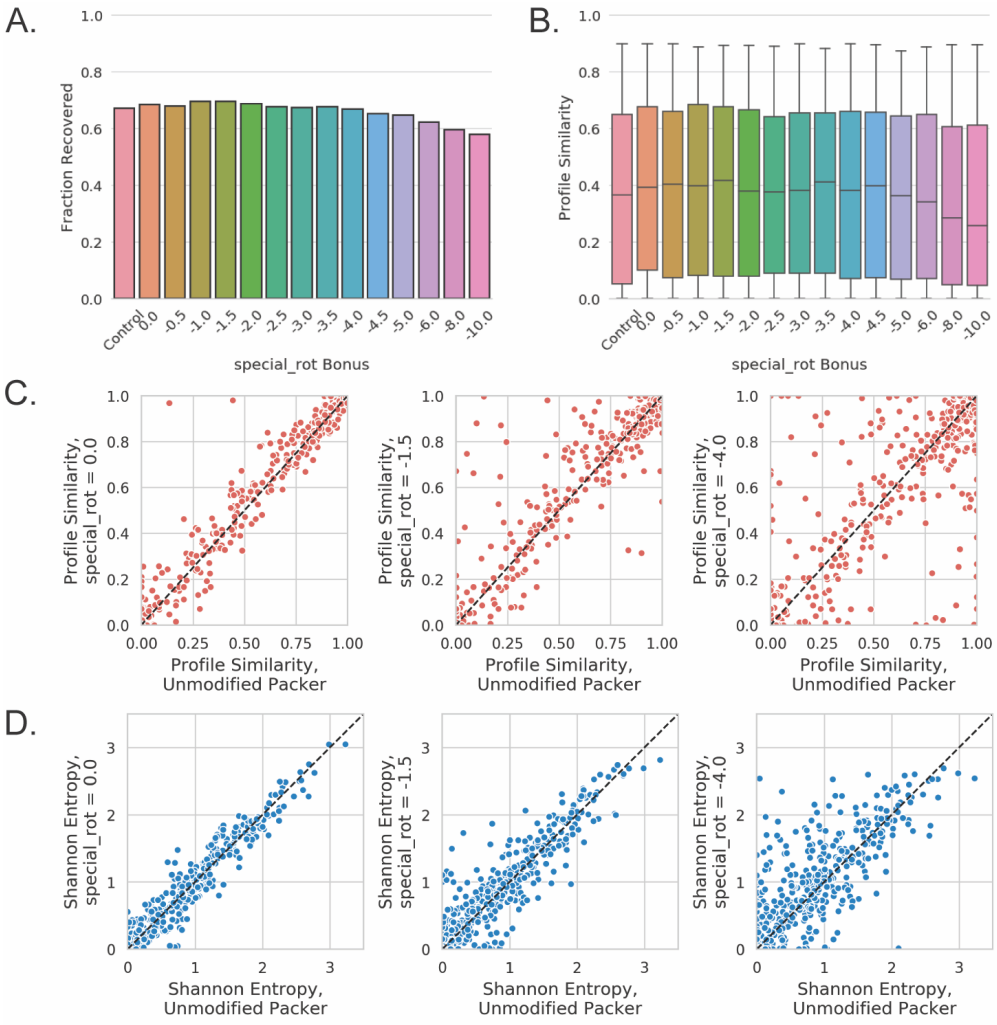


Figure 2.4: Complementary Rotamers Improve Binding Site Sequence Recovery. A. The fraction of designable positions in the benchmark complex set where the correct (wild-type) residue identity was incorporated at least 50% of the time (native sequence recovery) for a range of special_rot bonus values. Control indicates native sequence recovery when using the unmodified Packer. B. Box and whisker plots for profile similarities at designable positions in the benchmark set, where designable positions that achieve a profile similarity of $>$0.9 regardless of modifications to the Packer have been removed (plots with all positions are shown in Figure 2.8). C. Scatter plots showing profile similarity for individual designable positions at various special_rot bonus values (left: 0; middle: -1.5; right: -4.0) compared to the unmodified Packer. D. Scatter plots showing shannon entropy for individual designable positions at various special_rot bonus values (left: 0; middle: -1.5; right: -4.0) compared to the unmodified Packer.

improvement in profile similarity with the complementary rotamers as compared to the unmodified Packer (Figure 2.4 panel C, middle). It is important to note that the complementary RotamerSet with a special_rot bonus of -1.5 REU provided an improvement in profile similarity to design positions without diminishing high-similarity positions in both Packer conditions. On the contrary, for special_rot bonus values of -4.0 REU and below, profile similarities deteriorate for positions originally recovered by the unmodified Packer (Figure 2.4 panel C, right).

An additional benefit provided by complementary rotamers becomes evident when considering sequence entropy at designable positions: inclusion of complementary rotamers with a non-zero special_rot bonus leads to increased Shannon entropy for designable positions as compared to the unmodified Packer (Figure 2.4 panel D), while maintaining comparable if not better median profile similarities up to a value of -4.0 REU for the special_rot bonus (Figure 2.4 panel B). A key benefit of increased sequence entropy with complementary rotamers is demonstrated by the frequency and variety of polar and charged residues incorporated at several design positions (Figure 2.5). This behavior could lead to improvements over Rosetta's known propensity to incorporate small, hydrophobic residues over side chains capable of mediating hydrogen bonds, leading to frequent issues with buried unsatisfied hydrogen bonding donors and acceptors and poor sequence recovery in polar binding sites.[27]

2.2.5 Composite Binding Sites and Complementary Rotamers Applied to De Novo Design of Ligand Binders

Finally, we sought to demonstrate the combined application of the methods outlined in this work toward improved design of protein-ligand interfaces de novo. We selected 5 ligands from our match comparison set and selected matches for these ligands for design that passed designability filters based on ligand burial, composite binding site energy, and potential hydrogen bond satisfaction with the ligand (Table 2.9). We then designed the binding site environment of the selected matches using complementary rotamers. While a special_rot bonus of -1.5 was found to provide the greatest benefit in terms of native sequence recovery and profile similarity in the binding site recovery benchmark (Figure 2.4 panels A,B), we decided to also attempt complementary rotamers design with a special_rot bonus of -4.0. to investigate whether increased sequence entropy pro-

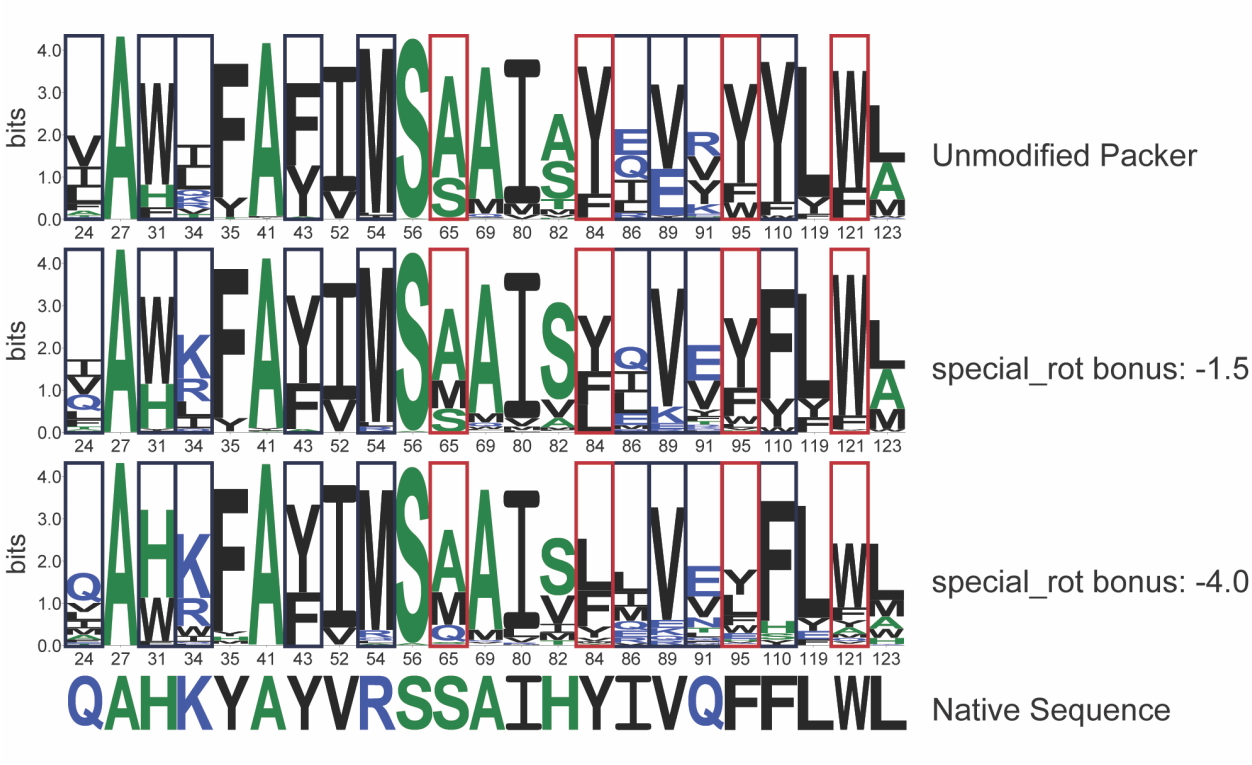


Figure 2.5: Sequence Logos for Recovery of Designable Binding Site Positions. Binding site recovery is shown for the engineered lipocalin DigA16 in complex with digitoxigenin (PDB ID: 1LNM, Ligand chemical component identifier: DTX). Sequence logos depict positional information content in bits, where residue identities are colored based on hydrophobicity (hydrophilic, blue; neutral, green; hydrophobic, black). Designable positions are in Rosetta numbering as in S6 Table. Design positions in navy blue boxes show improved native sequence recovery with the application of complementary rotamers. Specifically, positions with polar or charged residues show increased recovery with increased magnitude in the special_rot bonus. While the native residue identity of position 91 is not found, an isosteric residue is recovered with the application of complementary rotamers. Sequence recovery for positions in red boxes is negatively affected with increased special_rot bonus when compared to the unmodified Packer.

Table 2.2: Improved Design Metrics when Using Complementary Rotamers on Composite Binding Site Matches for 14 De Novo Ligand Binding Site Design Cases. Counts out of 14 cases of design on composite binding site matches where median metrics across 5000 designs that utilized composite RotamerSets with special_rot bonus values were improved as compared to the unmodified Packer. Ligand SASA and RosettaHoles are considered improved if the median decreased relative to the unmodified Packer. Shape complementarity is considered improved if the median increased relative to the unmodified packer.

special_rot Bonus	Ligand SASA	RosettaHoles	Shape Complementarity
0	6	9	10
-1.5	7	8	11
-3.0	12	9	11
-4.0	12	8	9

vided by a greater magnitude special_rot bonus (Figure 2.4 panel D) would benefit design (and the benchmark metrics remained comparable to the unmodified Packer up to this bonus). We also used a special_rot bonus of -3.0 as a midpoint between the aforementioned conditions.

A total of 5000 trajectories were attempted for each design condition and outputs were passed through commonly applied Rosetta filters to determine the quality of designs. In particular, the quality of the designs was judged based on shape complementarity of the binding site with the ligand,[31] packing within the binding site based on Rosetta's Holes[32] filter, and ligand solvent accessible surface area (SASA). These metrics were selected as tightly packed binding sites with high shape complementarity to the ligand were previously demonstrated to be essential for high ligand binding affinity and specificity in de novo designed proteins.[6]

The application of complementary rotamers improved key quality metrics when compared to design with the unmodified Packer for all special_rot bonus values attempted. Surprisingly, quality metrics improved even when the special_rot score bonus was set to zero (i.e. complementary rotamers were included but not favored beyond their Rosetta energies). This behavior is likely due to the increased degrees of sampling at the χ_1 and χ_2 torsions for complementary rotamers that were introduced to the Packer, resulting in finer sampling of rotamer conformations at the protein-ligand interface. The greatest benefit was imparted with a special_rot bonus of -3.0, where packing based on the Rosetta Holes score improved in 9/14 of design cases, ligand burial improved in 12/14 of design cases, and shape complementarity with the ligand improved in 11/14 of design

cases (Table 2.2).

When we focused on design metrics for individual composite binding site matches, we found that complementary rotamers provided a considerable benefit. For instance, we attempted design on a composite binding site for (E)-imidacloprid consisting of three leucines matched into the cavity of the TAP-p15 mRNA nuclear export factor (PDB ID: 1JKG). All metrics were improved compared to the unmodified packer when complementary rotamers with a special_rot bonus of -1.5 were applied (Figure 2.6). Importantly, complementary rotamers enriched for designs that possess high ($>$0.75) shape complementarity with imidacloprid as well as designs that reduced the amount of solvent accessible surface area on the ligand. Similarly, application of complementary rotamers with all special_rot bonus values enriched the number of designs with high shape complementarity and reduced solvent accessible surface area in a composite binding site match for naproxen (Figure 2.7). Here, a composite binding site composed of tryptophan, tyrosine, and phenylalanine was matched into the cavity of an NTF2-like protein (PDB ID: 2RCD). Importantly, for both the imidacloprid and naproxen binding site design cases, application of complementary rotamers did not merely improve one design metric at the expense of other metrics; complementary rotamers enriched for designs that were improved in all metrics as compared to designs generated by the unmodified Packer (Figure 2.6, Figure 2.7). When taken together, these results demonstrate that applying complementary rotamers to design of composite binding site matches generates designs for de novo ligand binding sites that exhibit favorable metrics that have been previously shown to be predictive of design success.

2.3 Discussion

In this work, we demonstrate new methods to improve the design of protein-ligand interfaces based on interactions observed in the PDB. By decomposing a small molecule into fragments, we leverage the wealth of structural information in the PDB to assemble pools of protein side chain interactions with fragments that compose the target ligand. We outline a new method to combine these discrete protein-fragment contacts into hundreds of thousands of composite binding sites for target ligands that are incorporated with high precision into scaffold proteins to nucleate a new binding site. We also outline a new method to bias the incorporation of side chain rotamers

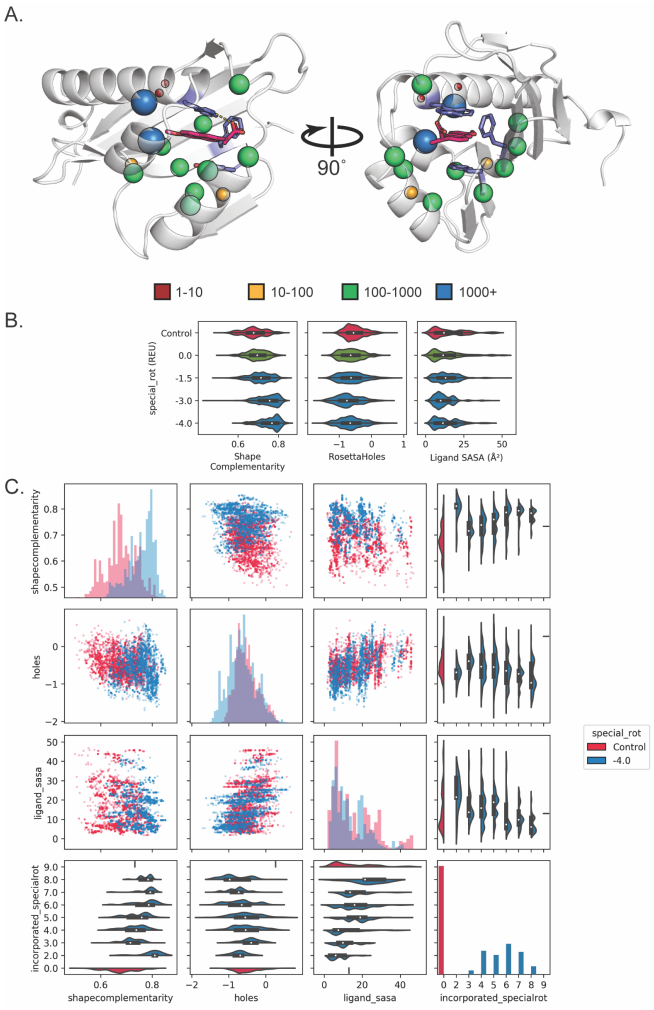


Figure 2.6: Composite binding sites combined with complementary rotamers improve design quality metrics for (E)-imidacloprid. A. Spheres indicate the number of complementary rotamers accepted per position to design the context of an all-leucine composite binding site (purple) for (E)-imidacloprid (hot pink) placed into the scaffold protein PDBID 1JKG. B. Violin plots depict the distribution of values obtained for shape complementarity, RosettaHoles, and ligands SASA metrics after design using complementary rotamers with special_rot bonus values, where Control designs were generated using the unmodified Packer. Application of complementary rotamers to design improves key metrics important for de novo design of ligand binding proteins with high specificity and selectivity. Application of complementary rotamers with the special_rot score bonus (blue) improves design metrics over merely adding complementary rotamers without the special_rot bonus (green) or the unmodified packer ("Control", red). Shape complementarity increases and RosettaHoles and ligand solvent accessible surface area (SASA) decrease. Design metrics improve as the special_rot bonus is increased from 0.0 to -4.0. C. Histograms on the diagonal show the design metric distribution for control designs as compared to designs with complementary rotamers and a special_rot bonus of -4.0 REU. Violin plots show the distributions of each design metrics as a function of incorporated complementary rotamers. Scatterplots show the correlation between different design metrics. Designs show simultaneous improvement in shape complementarity and binding site packing with the application of complementary rotamers.

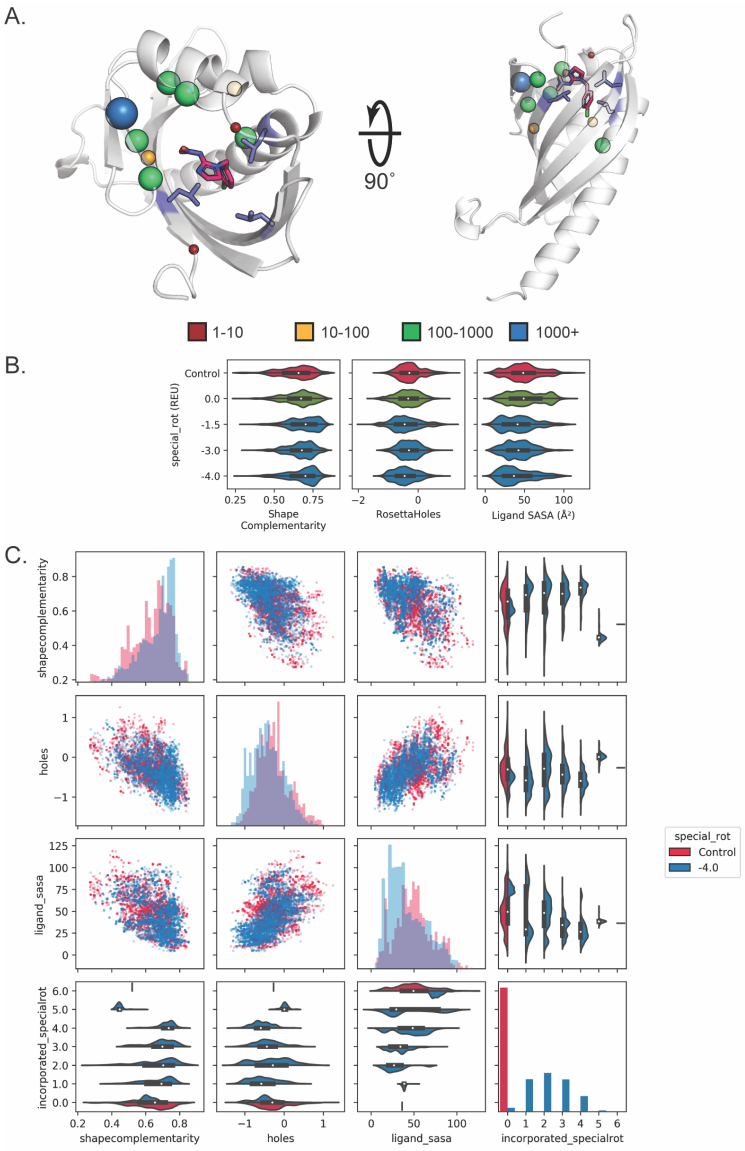


Figure 2.7: Composite binding sites combined with complementary rotamers improve design quality metrics for Naproxen. A. Spheres indicate the number of complementary rotamers accepted per position to design the context of a composite binding site composed of tryptophan, tyrosine, and phenylalanine (purple) for naproxen (hot pink) placed in the scaffold protein PDBID 2RCD. B. Plots and design metrics are as in Figure 2.6. Application of complementary rotamers improve shape complementarity and ligand SASA with increased special_rot bonus. C. Plots and design metrics are as in Figure 2.6. Designs show simultaneous improvement in shape complementarity, ligand SASA, and binding site packing with the application of complementary rotamers.

that reconstitute known interactions with ligand fragments during design. Finally, we demonstrate that the combination of these methods improves the quality of designs compared to Rosetta's conventional design machinery as determined by commonly used metrics.

The methods outlined in this work serve as a foundation for more complex strategies to design functional protein-ligand interactions. Protocols such as Rosetta's FastDesign, which incorporates a small amount of backbone change into the design process by iterating between fixed-backbone design and fixed-sequence minimization steps, could include the complementary rotamers described here to improve protein-ligand interface design. Other flexible design methods such as CoupledMoves[33] or Backrub Ensembles[34] may benefit from the application of complementary rotamers to further increase design sequence diversity and incorporation of key protein-ligand interactions in a binding site. Similarly, complementary rotamers could be incorporated into other side chain design approaches.[35,36,37]

While we were able to demonstrate effective application of protein-fragment interactions to the design of ligand binding sites, there are several classes of contacts that were not used productively. For instance, we noticed that 19.3% of fragment interactions in the contact pools generated for the ligands in this work were found to be mediated by backbone atoms (C, CA, O, N). This proportion agrees with the binding sites in the match comparison benchmark, where 17 of 93 contact constraints generated for existing complexes (see Methods) were mediated by a backbone atom. However, our composite binding site method focused on introducing productive side chain interactions with the ligand. In addition, complementary rotamers would not benefit from the addition of backbone interactions since they may not be introduced once the ligand is placed within a potential binding site context. It is also well known that water-mediated interactions are frequent in polar protein-small molecule interfaces[38,39] but are not considered here. Updating the method to better utilize these interactions could improve our ability to design functional protein-ligand interactions in the future.

Design with complementary rotamers relies on a fixed bonus provided by the special_rot score term to bias incorporation of observed protein-fragment interactions. While a special_rot bonus of -1.5 was found to provide the most benefit in native sequence recovery and profile similarity across all complexes in our benchmark set, the most adequate bonus to maximize these metrics varied between individual complexes. In addition, we found several design positions in the bind-

ing site recovery benchmark where the application of complementary rotamers was detrimental to sequence recovery (see examples in Figure 2.5). More tunable methods of applying bonuses for each complementary rotamer incorporated during design may provide a more generalizable benefit across varying design contexts. For instance, the design centric guidance terms implemented in Rosetta could provide a ramping score that provides a significant bonus for the first complementary rotamer incorporated, then ramp down the bonus for each additional rotamer introduced.[40] These metrics could also be implemented to bias incorporation of complementary rotamers that satisfy certain user-defined constraints such as hydrogen bonding with the ligand.

The methods described in this work provide a new framework for the de novo design of small molecule binding sites into proteins and directly address key shortcomings in current state-of-the-art methods for protein-ligand interface design. We drastically expand the possible combinations of protein-ligand contacts that can be incorporated into scaffold proteins by recombining observed protein-fragment contacts into composite binding sites, paving the way for the computational design of new binding sites for targets for which no co-complex structure exists in the PDB. To complement shortcomings in Rosetta's energy function with regards to ligands, we also apply contact pools to bias the design of protein-ligand interfaces toward contacts that are frequently observed to mediate interactions with defined ligand substructures in the PDB. When combined with Rosetta's success in de novo protein designed protein structures, we provide a foundation for the complete de novo design of functional ligand-binding proteins. For instance, where most existing methods attempt to build a binding site into an existing scaffold, just as we have demonstrated with three-residue composite binding sites and RosettaMatch, composite binding sites of arbitrary size can instead serve as a scaffold to build functionalized de novo proteins from the inside-out. We envision that composite binding sites and complementary rotamers will pioneer the design of proteins tailor built for specific functions and will augment our ability to design sensors and inducible transcription factors that require complex understanding of how protein-ligand interactions mediate chemistry and allostery.

2.4 Methods

2.4.1 Code Availability

The collection, processing, and application of contact pools derived from observed protein-ligand fragment interactions in the PDB is implemented as a Python3 package called BindingSites-FromFragments (BSFF). BSFF builds upon the Rosetta Molecular Modeling Suite,[5] PyRosetta (PyRosetta4.conda.linux.CentOS.python37.Release r240 2019.50+release.91b7a94),[41] ProDY,[42] and RDKit.[26] Each step of the protocol is accessible through a command-line interface with minimal intervention by the user after generating the initial inputs. BSFF creates and automatically populates a directory tree named with a three-letter code (typically the ligand's three character PDB identifier, referred to here as [/TGT]) in a user-defined location. All scripts and commands are included as part of the BSFF Repository (https://github.com/jaaamessszzz/ BindingSitesFrom-Fragments). Additional scripts for generating figures and submitting jobs are also available (https:// github.com/jaaamessszzz/BindingSitesFromFragments-Utilities).

2.4.2 Prepare Input Files

BSFF requires a Sybyl MOL2 representation of the target ligand as an initial input. For ligands with rotatable bonds, we used OpenEye Omega (version 2.5.1.4)[25] with default settings to generate a conformer library. A custom script (molfile_split_into_singles.py) uses Rosetta's molfile_to_params.py to generate a Rosetta params file for each ligand conformer. PDB representations of each conformer are also generated by molfile_split_into_singles.py so that unique atom names are consistent across all ligand conformers. These files are written to [/TGT/Inputs/Rosetta_Inputs].

2.4.3 Define Fragments

Avogadro[43] is used to generate user-defined fragments (a subset of atoms from the target ligand) derived from a single PDB representation produced in the previous step. This step relies on the user's chemical intuition to define fragments that represent local chemical substructures of the target ligand, which will be used to search the PDB to collect protein contacts with each

defined fragment. Each ligand fragment is saved as a PDB file and must conserve the unique atom names found in the full PDB representation of the target ligand. Typically, 5-10 fragments consisting of 3-10 atoms each are generated for a target ligand. The PDB representations of each fragment should be saved as Fragment_X.pdb to [/TGT/Inputs/Fragment_Inputs], where each fragment should be enumerated (replacing X in Fragment_X.pdb with fragment index).

2.4.4 Search for Protein-Fragment Interactions

PubChem substructure search[44] is used to identify small molecules that possess user-defined fragments as substructures. For each fragment, the PDB representation generated in the previous step is imported into the PubChem Sketcher Tool and converted into a SMILES string. Atom query attributes are defined to restrict search results to small molecules that possess user-defined fragments as substructures with the same connectivity and local chemical environment (e.g. bond order, aromaticity) as the target molecule. Explicit hydrogens are not removed before searching. A molecular weight cutoff of 800Da is applied to the search results. The top 1,000,000 search results are downloaded as a CSV file (cid, inchikey) and saved as Fragment_X.csv to [/TGT/Inputs/Fragment_Inputs], where the CSV name should correspond to the input fragment PDB. Complete SMILES search queries for each fragment are to be enumerated and saved in Fragment_Inputs.csv found in [/TGT/Inputs/Fragment_Inputs].

For each fragment, an intersection of PubChem search results with a complete list of ligands represented in the PDB from LigandExpo[45] yields all fragment-containing small molecules that may be found in complex with a protein in the PDB. For each fragment-containing molecule, the PDB REST API returns all structures with a 90% sequence identity cutoff where the small molecule is in complex with a protein (all complexes containing DNA/RNA are excluded). Ligands and protein-ligand complexes for each fragment are stored under [/TGT/PDB_search_results.json].

Alternatively, the Chemical Component Search function available through the PDB can be used to populate [/TGT/PDB_search_results.json]. Using this function may find compounds that are missed by PubChem, but at the expense of significant limitations in the ability to define fragment chemical environments in the search query.

2.4.5 Align

All protein-ligand complexes found in the previous step are processed to consolidate protein contacts in reference to each user-defined fragment. For each fragment-containing compound in a protein-ligand complex, the structure is checked to verify that 1) the ligand does not possess multiple occupancies and 2) all ligand atoms are resolved. If the structure passes these quality checks, the fragment substructure in the ligand is identified using SMILES representations and the Maximum Common Substructure search as implemented by RDKit. For each substructure contained within a ligand, all protein atoms within 12Å of the fragment substructure (including the fragment) are extracted and transformed such that the identified substructure is superimposed onto the user-generated PDB representation of the fragment. This process produces an aligned ensemble of all protein contacts with the defined fragment in the PDB that pass all quality filters.

2.4.6 Cluster

Agglomerative hierarchical clustering is performed for each fragment's ensemble of protein interactions to identify highly represented protein-fragment contacts. A feature vector is generated for residue in the ensemble consisting of: spatial position relative to the fragment, contact distance, interaction chemistry, and residue identity (Table 2.6). A filtering step removes hydrophobic residues (ACFGILMPVW) $>$4Å from any fragment atom or polar residues (DEHKNQRSTY) $>$3.5Å from any fragment atom. Residues are first clustered by protein-fragment interaction chemistry (Hamming distance, cutoff=2, linkage=complete) followed by spatial distribution about the fragment (Cosine distance, cutoff= $1-cos(20*\pi/180)$, linkage=average) to generate clusters of residues that mediate similar interaction types with specific parts of the target fragment.

2.4.7 Fragment Contact Analysis

Bootstrapping was applied to approximate the proportion of protein-fragment contact clusters (i.e. contact modes) that would be recovered given an observed fraction of the PDB. For a given fragment and sample size (ranging from 0% to 100% of the PDB at 5% intervals), 1000 samples were drawn from the PDB and the number of recovered clusters were recorded. Clusters were considered recovered if at least one contact in the cluster was sourced from a PDB in the sample. The

average fraction and standard deviation of clusters recovered were reported for each fragment and sample size and are shown in Figure 2.2.

2.4.8 Contact Pool Assembly

All clustered protein-fragment contacts are positioned relative to their source fragments in the target ligand to produce an ensemble of aligned protein contacts with the full target ligand. For ligand conformers, sidechains are transformed using three atoms each on the ligand and sidechain to maintain the contact geometry observed in the PDB. Additional filters are applied to remove redundant contacts and remove contacts from structures with missing (zero occupancy) protein atoms or alternative protein conformations (residues with ALTLOC records). Contacts are considered redundant if a sidechain has already been accepted with equivalent sidechain contact atoms, equivalent ligand contact atom, sidechain contact atoms RMSD $\leq$0.5Å, and Cα atom distance $\leq$1.5Å. Contact pools are limited to the best 5000 contacts as determined by Rosetta energy. Furthermore, the addition of hydrogen bond donors/acceptors (as determined by a negative hbond_sc score for the contact) to the contact pool are prioritized (up to 500 residues per donor/acceptor atom on the ligand) and remaining contacts are equally distributed about the ligand based on the ligand atom mediating contact with each residue. Alanine, glycine, proline, and cysteine residues were not added to contact pools in this work.

2.4.9 Assembly of Composite Binding Sites

For each target ligand conformer, a Markov Chain Monte Carlo protocol was applied to assemble discrete sidechains in the contact ensemble into composite binding sites. A composite binding site is initialized with three random sidechains from the contact ensemble and select two-body score terms in the Rosetta full-atom energy function (fa_rep, fa_atr, fa_elec, hbond_sc, fa_sol) are used to evaluate the overall energy of the binding site. During each move, a sidechain in the current binding site is replaced with a random sidechain from the contact ensemble. The new binding site is scored and the move is kept based on the Metropolis criterion, where the objective is to minimize the energy of the composite binding site. The temperature is ramped down geometrically across seven steps. Up to ten trajectories are performed for each contact ensemble and the lowest energy

100,000 solutions discovered per trajectory are recorded.

Selection of Binding Site Definitions for Existing Complexes. Binding site definitions were generated for existing protein-ligand complexes in the Match comparison benchmark. These constraints consist of the three lowest-energy contacts in the binding site for the given complex in terms of the two-body Rosetta energies used to assemble contact pools (fa_rep, fa_atr, fa_elec, hbond_sc, fa_sol) and the hbond_bb_sc term to account for binding site backbone contacts with the ligand. Existing protein-ligand complexes were relaxed with the Rosetta FastRelax protocol with the full REF2015 energy function and starting coordinate constraints (FastRelax command line option -relax:constrain_relax_to_start_coords) before determining constraint contacts.

2.4.10 Match Binding Sites into Scaffolds

RosettaMatch[14] is used to find binding site solutions that may be accommodated by existing backbone geometries in a protein scaffold set. A monomeric scaffold set was previously assembled.[6] RosettaMatch constraint files are automatically generated for the best 5,000 binding site solutions across all ligand conformers in terms of objective value (Rosetta energy). Constraint files are generated for the RosettaMatch algorithm using a custom script that calculates six degrees of freedom (1 distance, 2 angles, 3 torsions) for three ligand atoms and three side chain atoms that mediate the ligand-protein side chain interaction. Each angle and torsion degree of freedom is sampled at $\pm 5^{\circ}$ in addition to the ideal value, while the distance is fixed at the ideal value. Additional flags for the RosettaMatch application are as follows: -ex1 -ex2 -extrachi_cutoff 0 -bump_tolerance 0.5 -euclid_bin_size 1 -euler_bin_size 10 -match:consolidate_matches -match_grouper SameRotamerComboGrouper -output_matches_per_group 1.

Successful matches must pass additional filters where 1) the binding site energy as calculated in the context of the scaffold <0, 2) max fa_rep for any motif residue <25 Rosetta energy units (REU), and 3) max fa_sol for any motif residue <5 REU. A custom "hydrogen bond satisfaction" filter is also applied to ensure that positions exist in the match scaffold where a sidechain can be built to satisfy all potential hydrogen bond donor/acceptor atoms on the ligand. This filter generates rotamers (-ex1 -ex2 -extrachi_cutoff 0) for all positions within 10Å of the ligand. The filter reports which hydrogen bond donor/acceptor atoms on the ligand that are satisfied for all contacts found

where hbond_sc $<$ 0, fa_rep $<$ 10, and the sum of two-body terms used to solve for composite binding sites $<$ 10.

2.4.11 Selection of Binding Site Recovery Benchmark Set

Protein-ligand complexes for the binding site recovery benchmark were identified using the BindingMOAD[30] with the following search criteria: binding class, ligand mass of 200 - 800 Da, <1μM dissociation constant (K_D), and crystal structure of $<$2.5Å resolution. In addition to these criteria, complexes were further curated for binding sites composed of mainly side chain interactions, ligands with less than six rotatable bonds, and few waters and no metal coordination in the binding site. A set of 22 protein-ligand complexes was identified for the benchmark (Table 2.5). PDB ID 6M9B was used in place of BindingMOAD annotated streptavidin-biotin complexes.

2.4.12 Generation of Complementary Rotamers

Complementary rotamers are generated per-position in a protein-ligand interface to augment the set of rotamers Rosetta's Packer will use to design the binding site. Complementary rotamers are created in two steps. First, a new contact pool is assembled for the ligand placed into the scaffold in the protein-ligand complex. In addition to the aforementioned criteria for contact pool assembly, inverse rotamers[14] are generated for potential contacts in the context of the ligand binding site and are only added to the new contact pool if an inverse rotamer's mainchain (C, CA, N) RMSD to any position in the protein backbone is 2Å or less. This process yields a contact pool for interactions that may potentially be recapitulated within the context of the protein-ligand interface. Second, rotamers are generated for designable positions in the protein-ligand interface with extra χ-angle sampling at two half-step standard deviations for χ_1 and χ_2 (-packing:ex1:level 4 -packing:ex2:level 4) and one standard deviation for χ_3 and χ_4 (-packing:ex3:level 1 -packing:ex4:level 1). Rotamers that recapitulate contact geometries observed in the target ligand contact pool with RMSD $\leq$ 1.5Å are added to Rosetta's Packer RotamerSets. By default, up to 50 rotamers are added per residue identity to a position's RotamerSet, where each rotamer is flagged with the SPECIAL_ROT VariantType. If more than 50 rotamers are found to recapitulate contact pool interactions for a given position and residue identity, only the best rotamers by Rosetta energy are added to the

RotamerSet.

2.4.13 Design with Complementary Rotamers

Protein-ligand complexes (e.g. matches, or existing complexes from the PDB) are passed to conventional Rosetta design methods to redesign the binding site environment with the aid of complementary rotamers. The binding site environment is designed using fixed backbone packing as implemented in Rosetta[15] where the Packer RotamerSets are augmented with complementary rotamers flagged with the SPECIAL_ROT VariantType. All protein positions within 10Å of the ligand with Cα-Cβ vectors pointing toward the ligand (i.e. dot product of Cα-Cβ vector and Cα-centroid vectors > 0) are set to designable (i.e., allowed to change residue identity). In addition, these positions are passed to the ClashBasedShellSelector to identify a first shell of residues that are also set to designable. All designable positions are again passed to the ClashBasedShellSelector to identify a second shell of residues that are set to repackable (i.e., allowed to change conformation while keeping the residue identity the same). All other positions in the protein are fixed during design. All cysteines, glycines, and prolines in the protein are fixed. The following flags were used for design: -ex1 -ex2 -extrachi_cutoff 0 -use_input_sc -run:preserve_header -extra_res_fa params_path -total_threads 1, where params_path is the path to the params file for the ligand in the complex. Up to 10,000 independent design simulations are performed for each match with the special_rot score term added to the default REF2015 energy function to bias incorporation of complementary rotamers during design. For de novo binder design, matched binding site residues are set to repack only and constraints as defined during matching are applied using the AddOrRemoveMatchCsts mover to maintain contact geometries with the ligand during design. Rosetta filters were used to compute shape complementarity,[31] RosettaHoles score,[32] and ligand solvent accessible surface area metrics. The Binding Strain, Residue Interaction Energy, PackStat, and BuriedUnsatHbonds filters were also computed, but were less informative (S1 Appendix). An additional entry for chlorine (radius of 1.75Å) was added to the shape complementarity atom radius database file [database/scoring/score_functions/sc/sc_radii.lib] for ligands that contain chlorine atoms.

2.4.14 Profile Similarity and Sequence Recovery

For the binding site recovery benchmark, profile similarity was calculated at each design position as previously described,[33] where profile similarity is defined as 1 - Jensen-Shannon Divergence.[46] Jensen-Shannon Divergence was calculated per position for design sequences against the native complex sequence taken from the PDB. For sequence recovery, design positions were considered recovered if $>$50% of design sequences incorporated the residue identity observed in the native complex.

2.5 References

[1] Po Ssu Huang, Scott E. Boyken, and David Baker. “The coming of age of de novo protein design”. In: *Nature* 537.7620 (Sept. 2016), pp. 320–327. ISSN: 14764687. DOI: 10.1038/nature19946 (cit. on p. 3).

[2] Bettina Schreier et al. “Computational design of ligand binding is not a solved problem”. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.44 (Nov. 2009), pp. 18491–18496. ISSN: 00278424. DOI: 10.1073/pnas.0907950106 (cit. on p. 3).

[3] Jiayi Dou et al. “Sampling and energy evaluation challenges in ligand binding protein design”. In: *Protein Science* 26.12 (Dec. 2017), pp. 2426–2437. ISSN: 1469896X. DOI: 10.1002/pro.3317 (cit. on p. 3).

[4] David Baker. *What has de novo protein design taught us about protein folding and biophysics?* Apr. 2019. DOI: 10.1002/pro.3588 (cit. on p. 3).

[5] Andrew Leaver-Fay et al. “Rosetta3: An object-oriented software suite for the simulation and design of macromolecules”. In: *Methods in Enzymology* 487.C (2011), pp. 545–574. ISSN: 00766879. DOI: 10.1016/B978-0-12-381270-4.00019-6 (cit. on pp. 3, 24).

[6] Christine E. Tinberg et al. “Computational design of ligand-binding proteins with high affinity and selectivity”. In: *Nature* 501.7466 (Sept. 2013), pp. 212–216. ISSN: 00280836. DOI: 10.1038/nature12443 (cit. on pp. 3, 4, 12, 18, 28).

[7] Matthew J. Bick et al. “Computational design of environmental sensors for the potent opioid fentanyl”. In: *eLife* 6 (Sept. 2017). ISSN: 2050084X. DOI: 10.7554/eLife.28909 (cit. on p. 3).

[8] Anum A. Glasgow et al. “Computational design of a modular protein sense-response system”. In: *Science* 366.6468 (Nov. 2019), pp. 1024–1028. ISSN: 10959203. DOI: 10.1126/science.aax8780 (cit. on pp. 3, 4).

[9] Jiayi Dou et al. “De novo design of a fluorescence-activating β-barrel”. In: *Nature* 561.7724 (Sept. 2018), pp. 485–491. ISSN: 14764687. DOI: 10.1038/s41586-018-0509-0 (cit. on pp. 3, 5).

[10] Angela Lombardi et al. “De novo design of four-helix bundle metalloproteins: One scaffold, diverse reactivities”. In: *Accounts of Chemical Research* 52.5 (Apr. 2019), pp. 1148–1159. ISSN: 15204898. DOI: 10.1021/acs.accounts.8b00674 (cit. on p. 4).

[11] Nicholas F. Polizzi et al. “De novo design of a hyperstable non-natural protein-ligand complex with sub-Å accuracy”. In: *Nature Chemistry* 9.12 (Dec. 2017), pp. 1157–1164. ISSN: 17554349. DOI: 10.1038/nchem.2846 (cit. on p. 4).

[12] Ulrike Scheib et al. “Change in protein-ligand specificity through binding pocket grafting”. In: *Journal of Structural Biology* 185.2 (Feb. 2014), pp. 186–192. ISSN: 10478477. DOI: 10.1016/j.jsb.2013.06.002 (cit. on p. 4).

[13] Wei Yang and Luhua Lai. “Computational design of ligand-binding proteins”. In: *Current Opinion in Structural Biology* 45 (Aug. 2017), pp. 67–73. ISSN: 1879033X. DOI: 10.1016/j.sbi.2016.11.021 (cit. on p. 4).

[14] Alexandre Zanghellini et al. “New algorithms and an in silico benchmark for computational enzyme design”. In: *Protein Science* 15.12 (Dec. 2006), pp. 2785–2794. ISSN: 09618368. DOI: 10.1110/ps.062353106 (cit. on pp. 4, 28, 29).

[15] Brian Kuhlman and David Baker. “Native protein sequences are close to optimal for their structures”. In: *Proceedings of the National Academy of Sciences of the United States of America* 97.19 (Sept. 2000), pp. 10383–10388. ISSN: 00278424. DOI: 10.1073/pnas.97.19.10383 (cit. on pp. 5, 13, 30).

[16] Maxim V. Shapovalov and Roland L. Dunbrack. “A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions”. In: *Structure* 19.6 (June 2011), pp. 844–858. ISSN: 09692126. DOI: 10.1016/j.str.2011.03.019 (cit. on pp. 5, 13).

[17] Rebecca F. Alford et al. “The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design”. In: *Journal of Chemical Theory and Computation* 13.6 (June 2017), pp. 3031–3048. ISSN: 15499626. DOI: 10.1021/acs.jctc.7b00125 (cit. on pp. 5, 10).

[18] Hahnbeom Park et al. “Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules”. In: *Journal of Chemical Theory and*

Computation 12.12 (Dec. 2016), pp. 6201–6212. ISSN: 15499626. DOI: 10.1021/acs.jctc.6b00819 (cit. on p. 5).

[19] Summer B. Thyme, David Baker, and Philip Bradley. “Improved modeling of side-chain-base interactions and plasticity in protein-dna interface design”. In: *Journal of Molecular Biology* 419.3-4 (June 2012), pp. 255–274. ISSN: 00222836. DOI: 10.1016/j.jmb.2012.03.005 (cit. on pp. 7, 13).

[20] Jörg Degen et al. “On the art of compiling and using ’drug-like’ chemical fragment spaces”. In: *ChemMedChem* 3.10 (Oct. 2008), pp. 1503–1507. ISSN: 18607179. DOI: 10.1002/cmdc.200800178 (cit. on p. 7).

[21] Xiao Qing Lewell et al. “RECAP - Retrosynthetic Combinatorial Analysis Procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry”. In: *Journal of Chemical Information and Computer Sciences* 38.3 (1998), pp. 511–522. ISSN: 00952338. DOI: 10.1021/ci970429i (cit. on p. 7).

[22] Tanja Kortemme, Alexandre V Morozov, and David Baker. “An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes.” In: *Journal of molecular biology* 326.4 (Feb. 2003), pp. 1239–59. ISSN: 0022-2836. DOI: 10.1016/s0022-2836(03)00021-4 (cit. on p. 10).

[23] Themis Lazaridis and Martin Karplus. “Effective energy function for proteins in solution”. In: *Proteins: Structure, Function, and Bioinformatics* 35.2 (May 1999), pp. 133–152. ISSN: 1097-0134. DOI: 10.1002/(SICI)1097-0134(19990501)35:2<133::AID-PROT1>3.0.CO;2-N (cit. on p. 10).

[24] Warren L DeLano. “Unraveling hot spots in binding interfaces: progress and challenges”. In: *Current opinion in structural biology* 12.1 (2002), pp. 14–20 (cit. on p. 10).

[25] Paul C.D. Hawkins et al. “Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and cambridge structural database”. In: *Journal of Chemical Information and Modeling* 50.4 (Apr. 2010), pp. 572–584. ISSN: 15499596. DOI: 10.1021/ci100031x (cit. on pp. 10, 24).

[26] *RDKit: Open-source cheminformatics* (cit. on pp. 10, 24).

[27] Brittany Allison et al. “Computational design of protein-small molecule interfaces”. In: *Journal of Structural Biology* 185.2 (Feb. 2014), pp. 193–202. ISSN: 10478477. DOI: 10.1016/j.jsb.2013.08.003 (cit. on pp. 13, 16).

[28] Samuel Deluca, Brent Dorr, and Jens Meiler. “Design of native-like proteins through an exposure-dependent environment potential”. In: *Biochemistry* 50.40 (Oct. 2011), pp. 8521–8528. ISSN: 15204995. DOI: 10.1021/bi200664b (cit. on p. 13).

[29] Matthew J. O’Meara et al. “Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta”. In: *Journal of Chemical Theory and Computation* 11.2 (Feb. 2015), pp. 609–622. ISSN: 15499626. DOI: 10.1021/ct500864r (cit. on p. 13).

[30] Liegi Hu et al. “Binding MOAD (Mother of All Databases)”. In: *Proteins: Structure, Function and Genetics* 60.3 (Aug. 2005), pp. 333–340. ISSN: 08873585. DOI: 10.1002/prot.20512 (cit. on pp. 13, 29).

[31] Michael C. Lawrence and Peter M. Colman. “Shape complementarity at protein/protein interfaces”. In: *Journal of Molecular Biology* 234.4 (Dec. 1993), pp. 946–950. ISSN: 00222836. DOI: 10.1006/jmbi.1993.1648 (cit. on pp. 18, 30).

[32] Will Sheffler and David Baker. “RosettaHoles: Rapid assessment of protein core packing for structure prediction, refinement, design, and validation”. In: *Protein Science* 18.1 (Jan. 2009), pp. 229–239. ISSN: 09618368. DOI: 10.1002/pro.8 (cit. on pp. 18, 30).

[33] Noah Ollikainen, René M. de Jong, and Tanja Kortemme. “Coupling Protein Side-Chain and Backbone Flexibility Improves the Re-design of Protein-Ligand Specificity”. In: *PLoS Computational Biology* 11.9 (2015). ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004335 (cit. on pp. 22, 31).

[34] Noah Ollikainen et al. “Flexible backbone sampling methods to model and design protein alternative conformations”. In: *Methods in Enzymology* 523 (Jan. 2013), pp. 61–85. ISSN: 15577988. DOI: 10.1016/B978-0-12-394292-0.00004-7 (cit. on p. 22).

[35] Christoph Malisi et al. “Binding Pocket Optimization by Computational Protein Design”. In: *PLoS ONE* 7.12 (Dec. 2012). ISSN: 19326203. DOI: 10.1371/journal.pone.0052505 (cit. on p. 22).

[36] Pablo Gainza et al. “Osprey: Protein design with ensembles, flexibility, and provable algorithms”. In: *Methods in Enzymology* 523 (Jan. 2013), pp. 87–107. ISSN: 15577988. DOI: 10.1016/B978-0-12-394292-0.00005-9 (cit. on p. 22).

[37] Mark A. Hallen, Daniel A. Keedy, and Bruce R. Donald. “Dead-end elimination with perturbations (DEEPer): A provable protein design algorithm with continuous sidechain and backbone flexibility”. In: *Proteins: Structure, Function and Bioinformatics* 81.1 (Jan. 2013), pp. 18–39. ISSN: 08873585. DOI: 10.1002/prot.24150 (cit. on p. 22).

[38] John E. Ladbury. “Just add water! The effect of water on the specificity of protein- ligand binding sites and its potential application to drug design”. In: *Chemistry and Biology* 3.12 (Dec. 1996), pp. 973–980. ISSN: 10745521. DOI: 10.1016/S1074-5521(96)90164-7 (cit. on p. 22).

[39] Benjamin Breiten et al. “Water networks contribute to enthalpy/entropy compensation in protein-ligand binding”. In: *Journal of the American Chemical Society* 135.41 (Oct. 2013), pp. 15579–15584. ISSN: 00027863. DOI: 10.1021/ja4075776 (cit. on p. 22).

[40] Parisa Hosseinzadeh et al. “Comprehensive computational design of ordered peptide macrocycles”. In: *Science* 358.6369 (Dec. 2017), pp. 1461–1466. ISSN: 10959203. DOI: 10.1126/science.aap7577 (cit. on p. 23).

[41] Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J. Gray. “PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta”. In: *Bioinformatics* 26.5 (Mar. 2010), pp. 689–691. DOI: 10.1093/BIOINFORMATICS (cit. on p. 24).

[42] Ahmet Bakan, Lidio M. Meireles, and Ivet Bahar. “ProDy: Protein Dynamics Inferred from Theory and Experiments”. In: *Bioinformatics* 27.11 (June 2011), pp. 1575–1577. DOI: 10.1093/BIOINFORMATICS (cit. on p. 24).

[43] Marcus D. Hanwell et al. “Avogadro: An advanced semantic chemical editor, visualization, and analysis platform”. In: *Journal of Cheminformatics* 4.8 (Aug. 2012), p. 17. ISSN: 17582946. DOI: 10.1186/1758-2946-4-17 (cit. on p. 24).

[44] Sunghwan Kim et al. “PubChem 2019 update: improved access to chemical data”. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D1102–D1109. DOI: 10.1093/nar/gky1033 (cit. on p. 25).

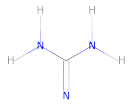
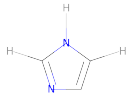
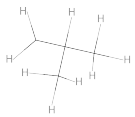
[45] Zukang Feng et al. “Ligand Depot: A data warehouse for ligands bound to macromolecules”. In: *Bioinformatics* 20.13 (Sept. 2004), pp. 2153–2155. ISSN: 13674803. DOI: 10.1093/bioinformatics/bth214 (cit. on p. 25).

[46] Golan Yona and Michael Levitt. “Within the twilight zone: A sensitive profile-profile comparison tool based on information theory”. In: *Journal of Molecular Biology* 315.5 (Feb. 2002), pp. 1257–1275. ISSN: 00222836. DOI: 10.1006/jmbi.2001.5293 (cit. on p. 31).

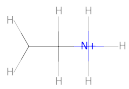
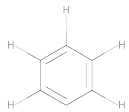
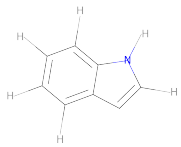
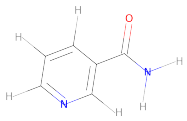
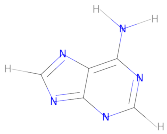
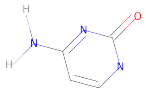
2.6 Supplemental

2.6.1 Tables

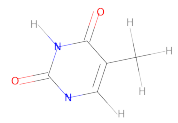
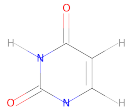
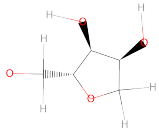
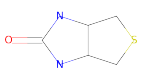
Table 2.3: Statistics for Fragments Used to Investigate Contact Diversity. Kekule structures and corresponding SMILES strings are provided for all fragments used to investigate the availability of protein-fragment contact information in the PDB. The total number of unique residue-fragment contacts and the number of unique clusters (i.e. contact modes) represented by these residue-fragment contacts are reported.

Fragment	SMILES	Total Residues	Total Clusters	Fraction PDB Sampled for >80% Contact Recovery
	[N;v3,v4]=C([N;v3]([H])[H])[N;v3]([H])[H]	3699	467	0.45
	CC(=O)N([H])[H]	19077	932	0.35
	CC(=O)O[H]	64489	1699	0.20
	C1=C([N](C(=N1)[H])[H])[H]	3722	592	0.40
	C([H])([H])C([H])(C([H])([H])[H])C([H])([H])[H]	6432	492	0.40

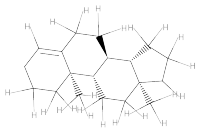
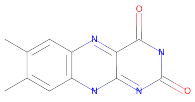
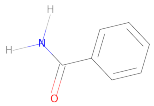
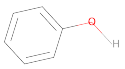
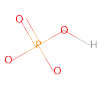
2.3 Continued

Fragment	SMILES	Total Residues	Total Clusters	Fraction PDB Sampled for >80% Contact Recovery
	C([H])([H])C([N+]([H])([H])[H])([H])[H]	2814	442	0.50
	C1=C(C(=C(C(=C1[H])[H])[H])[H])[H]	10063	746	0.35
	C1=C([N]([H])C2=C([H])C(=C(C(=C12)[H])[H])[H])[H]	2089	314	0.55
	N1=C(C(=C(C(=C1[H])C(=O)N([H])[H])[H])[H])[H]	3414	319	0.45
	N2=C1C(=C(N([H])[H])N=C([N]1)[H])N=C2[H]	68520	1046	0.25
	N1C=CC(=NC1=O)N([H])[H]	2123	282	0.50

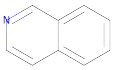
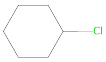
2.3 Continued

Fragment	SMILES	Total Residues	Total Clusters	Fraction PDB Sampled for >80% Contact Recovery
	N1C(=O)N(C(=O)C(=C1[H])C([H])([H])[H])[H]	2116	228	0.50
	[N]1C(=NC2=C1N=C(N(C2=O)[H])N([H])[H])[H]	12387	580	0.40
	N1C(N(C(C(=C1[H])[H])=O)[H])=O	6040	440	0.45
	[C]1([C@H](O[H])[C@@H]([C@H](O1)C(O)([H])[H])O[H])([H])[H]	11621	754	0.35
	C1=NC=CS1	1573	215	0.55
	C1SCC2NC(=O)NC12	1051	34	0.20

2.3 Continued

Fragment	SMILES	Total Residues	Total Clusters	Fraction PDB Sampled for >80% Contact Recovery
	C1(C(CC(=C2C(C([C@@H]3[C@@H]([C@@]12C([H])([H])[H])C(C([C@]4([C@H]3C(C(C4[H])([H])[H])([H])[H])C([H])([H])[H])([H])[H])([H])[H])([H])[H])([H])[H])([H])[H])([H])[H])([H])[H]	511	102	0.55
	N3=C2C(=NC1=C(C=C(C(=C1)C)C)N2)C(NC3=O)=O	21494	685	0.30
	[C]1=CC(=[C][C]=[C]1)C(=O)N([H])[H]	284	84	0.65
	C1=CC=C(C=C1)O[H]	12964	896	0.35
	N1=N[N]C=N1	166	52	0.65
	O=[P](O)(O[H])O	78948	1371	0.15

2.3 Continued

Fragment	SMILES	Total Residues	Total Clusters	Fraction PDB Sampled for >80% Contact Recovery
	[S](=O)(=O)(N)C	1631	252	0.45
	C1=C2C(=CC=C1)C=NC=C2	533	126	0.60
	C1=CC(=CC=C1)[Cl]	7592	599	0.40
	C1=CN=CC=C1	17058	755	0.35
	[C]1[C][C][N][C][C]1	18883	811	0.35
	C1=CN=CN=C1	52449	995	0.25

2.3 Continued

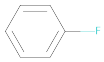
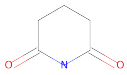
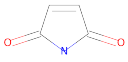
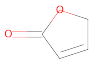
Fragment	SMILES	Total Residues	Total Clusters	Fraction PDB Sampled for >80% Contact Recovery
	N1[C][C]N[C][C]1	3799	551	0.45
	FC1=CC=CC=C1	4243	523	0.45
	O=C1NC(=O)CC[C]1	170	26	0.65
	C1(C=CC(N1)=O)=O	485	82	0.60
	C1=CCOC1=O	357	96	0.65

Table 2.4: Application Ligands. Full name and chemical component identifier for ligands used for the Match comparison benchmark and forward design with matched composited binding sites. All existing PDBs where a protein creates a contact interface with the ligand are provided.

Ligand	Chemical Component Identifier	Complex PDB(s)
(5Z)-5-(3,5-difluoro-4-hydroxybenzylidene)-2,3-dimethyl-3,5-dihydro-4H-imidazol-4-one	38E	6CZH, 6CZI
8,9-DIHYDRO-9-HYDROXY-AFLATOXIN B1	AFN	None.
ATRAZINE	ATZ	5PRC
DIGOXIGENIN	DOG	1LKE, 3RA7, 4J8T, 5BVB
IBUPROFEN	IBP	6U4X, 5JQB, 4RS0, 4PH9, 4JTR, 3P6H, 3IB2, 2WD9, 2PWS, 2BXG, 1EQG
(E)-imidacloprid	IM4	3WTH, 3C79
LUMIFLAVIN	LFN	6ASL, 2CCC
Naproxen	NPS	4ZBR, 4OR0, 4PO0, 4JQ1, 4FJP, 3R58, 3NT1, 2VDB

Table 2.5: BindingMOAD Complexes used for Binding Site Sequence Recovery. PDB ID, PDB description, ligand chemical component identifier, and full ligand name for BindingMOAD protein-ligand complexes applied in the binding site recovery benchmark.

PDB	PDB Description	Ligand	Ligand Name
6M9B	Wild-type streptavidin in complex with biotin solved by native SAD with data collected at 6 keV	BTN	Biotin
5T52	LECTIN FROM BAUHINIA FORFICATA IN COMPLEX WITH GALNAC	NGA	N-ACETYL-D-GALACTOSAMINE
1LNM	ANTICALIN DIGA16 IN COMPLEX WITH DIGITOXIGENIN	DTX	DIGITOXIGENIN
5HZ6	FABP4 in complex with 6-Chloro-2-isopropyl-4-(3-isopropyl-phenyl)-quinoline-3-carboxylic acid	65Y	6-Chloro-2-isopropyl-4-(3-isopropyl-phenyl)-quinoline-3-carboxylic acid
5HZ8	FABP4_3 in complex with 6,8-Dichloro-4-phenyl-2-piperidin-1-yl-quinoline-3-carboxylic acid	65Z	6,8-dichloro-4-phenyl-2-(piperidin-1-yl)quinoline-3-carboxylic acid
1LKE	ENGINEERED LIPOCALIN DIGA16 IN COMPLEX WITH DIGOXIGENIN	DOG	DIGOXIGENIN
1N0S	ENGINEERED LIPOCALIN FLUA IN COMPLEX WITH FLUORESCEIN	FLU	2-(6-HYDROXY-3-OXO-3H-XANTHEN-9-YL)-BENZOIC ACID
3OKI	Crystal structure of human FXR in complex with (2S)-2-[2-(4-chlorophenyl)-1H-benzimidazol-1-yl]-N,2-dicyclohexylethanamide	OKI	(2S)-2-[2-(4-chlorophenyl)-1H-benzimidazol-1-yl]-N,2-dicyclohexylethanamide
5EDB	human fatty acid binding protein 4 in complex with 6-Chloro-2-methyl-4-phenyl-quinoline-3-carboxylic acid at 1.18A	5M8	6-chloranyl-2-methyl-4-phenyl-quinoline-3-carboxylic acid
5URA	Enantiomer-Specific Binding of the Potent Antinociceptive Agent SBFI-26 to Anandamide transporters FABP7	8KS	(1S,2S,3S,4S)-3-[(naphthalen-1-yl)oxy]carbonyl-2,4-diphenylcyclobutane-1-carboxylic acid
1SRI	STRUCTURE-BASED DESIGN OF SYNTHETIC AZOBENZENE LIGANDS FOR STREPTAVIDIN	DMB	2-((3',5'-DIMETHYL-4'-HYDROXYPHENYL)AZO)BENZOIC ACID
1TOU	Crystal structure of human adipocyte fatty acid binding protein in complex with a non-covalent ligand	B1V	2-[(2-OXO-2-PIPERIDIN-1-YLETHYL)SULFANYL]-6-(TRIFLUOROMETHYL)PYRIMIDIN-4-OL
2IZL	STREPTAVIDIN-2-IMINOBIOTIN PH 7.3 I222 COMPLEX	IMI	2-IMINOBIOTIN
4AFH	Capitella teleta AChBP in complex with lobeline	L0B	Alpha-Lobeline
4QAC	X-RAY STRUCTURE OF ACETYLCHOLINE BINDING PROTEIN (ACHBP) IN COMPLEX WITH 4-(4-methylpiperidin-1-yl)-6-(4-(trifluoromethyl)phenyl)pyrimidin-2-amine	KK3	4-(4-methylpiperidin-1-yl)-6-[4-(trifluoromethyl)phenyl]pyrimidin-2-amine
2QRY	Periplasmic thiamin binding protein	TPS	THIAMIN PHOSPHATE
4AFG	Capitella teleta AChBP in complex with varenicline	QMR	VARENICLINE
4B5D	Capitella teleta AChBP in complex with psychonicline (3-((2(S)-Azetidinyl)methoxy)-5-((1S,2R)-2-(2-hydroxyethyl)cyclopropyl)pyridine)	SW4	2-[(1R,2S)-2-[5-[[(2S)-azetidin-2-yl]methoxy]pyridin-3-yl]cyclopropyl]ethanol
5J5G	X-Ray Crystal Structure of Acetylcholine Binding Protein (AChBP) in Complex with 6-(4-methoxyphenyl)-N4,N4-bis[(pyridin-2-yl)methyl]pyrimidine-2,4-diamine	6GF	6-(4-methoxyphenyl)-N 4 ,N 4 -bis[(pyridin-2-yl)methyl]pyrimidine-2,4-diamine
3CZ1	Dimeric crystal structure of a pheromone binding protein from Apis mellifera in complex with the n-butyl benzene sulfonamide at pH 7.0	NBB	N-BUTYL-BENZENESULFONAMIDE
2XN3	Crystal structure of thyroxine-binding globulin complexed with mefenamic acid	ID8	2-[(2,3-DIMETHYLPHENYL)AMINO]BENZOIC ACID
5J5I	X-Ray Crystal Structure of Acetylcholine Binding Protein (AChBP) in Complex with 4-(2-amino-6-bis[(pyridin-2-yl)methyl]aminopyrimidin-4-yl)phenol	6GM	4-(2-amino-6-bis[(pyridin-2-yl)methyl]aminopyrimidin-4-yl)phenol

Table 2.6: Feature Vector Components. Descriptions of individual feature vector components used to cluster residue-fragment interactions into unique contact modes using hierarchical agglomerative clustering.

Value	Description
Angstroms	X component, vector from fragment centroid to closest residue atom
Angstroms	Y component, vector from fragment centroid to closest residue atom
Angstroms	Z component, vector from fragment centroid to closest residue atom
0 | 1	Sidechain contact (1) OR backbone contact (0)
0 | 1	Ligand Polar Contact (1) OR Ligand Non-polar Contact (0)
0 | 1	Side chain possesses (1) or does not possess (0) hydrogen bond donor/acceptor (DEHKNQRSTY)
0 | 1	Hydrophobic, aliphatic (AILV) contact (1) or not (0)
0 | 1	Hydrophobic, aromatic (FWY) contact (1) or not (0)
0 | 1	Polar (NCQMST) contact (1) or not (0)
0 | 1	Charged, Acidic (DE) contact (1) or not (0)
0 | 1	Charged, Basic (HKR) contact (1) or not (0)
0 | 1	Glycine contact (1) or not (0)
0 | 1	Proline contact (1) or not (0)
0 | 1	Backbone carbonyl contact (1) or not (0)
0 | 1	Backbone amino contact (1) or not (0)
0 | 1	Backbone C/CA contact (1) or not (0)

Table 2.7: Counts for Improved Metrics Across All Attempted Designs. Counts out of 14 composite binding site matches where median metrics across 5000 designs that utilized composite RotamerSets with special_rot bonuses were improved or impared as compared to designs generated with unmodified Packer. Ligand SASA, binding strain (bindingstrain), PackStat, residue interaction energy (residueie), and shape complementarity are considered improved/impaired if the median increased/decreased 10% relative to the unmodified Packer. RosettaHoles is considered improved/impared if the median increased/decreased by 0.25 units relative to the unmodified packer. Number of hydrogen bonds made with the ligand (hbonds) and heavy buried unsatisfied hydrogen bond donor/acceptor count (heavyburiedunsats) is considered improved/impaired if the median count is more/less than the median count for the unmodified Packer.

special_rot bonus	bindingstrain	hbonds	heavyburiedunsats	holes	packstat	residueie	ligand_sasa	shapecomplementarity
0	7	1	2	9	11	8	6	10
-1.5	6	1	1	8	7	7	7	11
-3.0	2	3	0	9	8	8	12	11
-4.0	2	3	2	8	8	5	12	9

Table 2.8: Binding Site Benchmark Design Details. Statistics for complementary rotamers generated and applied to the binding site recovery benchmark. Rosetta numbering and the corresponding PDB numbering for scaffold PDBs are provided. Here we report the number of designable positions as well as the number of complementary rotamers generated, accepted (i.e. passed Rosetta energy and RMSD filters, see Methods), and applied to design for each benchmark binding site. Only the best 50 rotamers per residue type per position were applied top design.

Benchmark Complex	Designable Positions (Rosetta numbering)	Designable Positions (PDB numbering)	Total rotamers, accepted	Total rotamers, applied	Total rotamers, generated	Designable positions, count
5edb-clean	13, 16, 19, 20, 22, 23, 24, 26, 28, 30, 33, 36, 39, 56, 58, 60, 61, 63, 77, 78, 79, 80, 107, 127, 129, 131	11A ,14A ,17A ,20A ,21A ,24A ,26A ,34A ,37A ,41A ,43A ,52A ,54A ,56A ,58A ,59A ,61A ,75A ,76A ,77A ,105A ,116A ,125A ,127A ,129A	7972	1544	123299	26
5HZ6-clean	10, 13, 16, 19, 20, 23, 25, 29, 30, 32, 33, 36, 40, 41, 51, 53, 54, 55, 57, 58, 59, 60, 74, 75, 76, 77, 124, 126, 128	11A ,14A ,17A ,20A ,21A ,26A ,30A ,31A ,33A ,34A ,37A ,41A ,43A ,52A ,54A ,55A ,56A ,58A ,59A ,60A ,61A ,75A ,76A ,77A ,114A ,116A ,125A ,127A ,129A	9768	2024	166273	29
5HZ8-clean	14, 17, 20, 23, 24, 29, 34, 37, 40, 44, 57, 59, 61, 62, 64, 78, 79, 80, 81, 108, 110 ,119, 121, 130, 132	11A ,14A ,17A ,20A ,21A ,26A ,31A ,34A ,37A ,38A ,41A ,54A ,56A ,58A ,59A ,61A ,75A ,76A ,77A ,105A ,107A ,114A ,116A ,118A ,127A ,129A	10876	1901	110344	25
5J5G-clean	151, 152, 193, 200, 253, 274, 276, 277, 278, 279, 296, 324, 325, 332, 333, 334, 335, 385	143A ,144A ,146A ,148A ,185A ,190A ,192A ,32B ,34B ,36B ,53B ,55B ,56B ,57B ,58B ,73B ,81B ,101B ,103B ,104B ,106B ,111B ,112B ,113B ,114B ,164B	3798	471	47205	18
5J5I-clean	42, 61, 63, 65, 66, 111, 112, 119, 120, 121, 122, 172, 364, 365, 367, 404, 405, 406, 411, 413	-6A ,32A ,34A ,36A ,53A ,55A ,57A ,58A ,73A ,78A ,81A ,99A ,101A ,103A ,104A ,106A ,111A ,112A ,113A ,114A ,164A ,143E ,144E ,146E ,183E ,184E ,185E ,190E ,192E	2720	604	62164	20
5ura-clean	13, 16, 19, 20, 22, 23, 24, 28, 32, 33, 35, 37, 38, 39, 56, 58, 60, 61, 63, 77, 78, 79, 80, 118, 120, 129, 131	11A ,14A ,17A ,20A ,21A ,26A ,33A ,35A ,37A ,41A ,54A ,56A ,58A ,59A ,61A ,75A ,76A ,77A ,105A ,116A ,118A ,127A ,129A	6594	1636	130454	27
1TOU-clean	10, 13, 16, 29, 33, 36, 37, 39, 52, 53, 54, 55, 57, 58, 59, 60, 61, 74, 75, 76, 77, 93, 104, 115, 126	9A ,10A ,13A ,16A ,19A ,20A ,25A ,32A ,33A ,36A ,37A ,39A ,40A ,41A ,42A ,51A ,53A ,54A ,55A ,57A ,58A ,59A ,60A ,74A ,75A ,76A ,113A ,115A ,124A ,126A ,128A	413	177	51463	25

2.8 Continued

Benchmark Complex	Designable Positions (Rosetta numbering)	Designable Positions (PDB numbering)	Total rotamers, accepted	Total rotamers, applied	Total rotamers, generated	Designable positions, count
6m9b-clean	9, 10, 11, 13, 15, 29, 30, 31, 32, 33, 35, 36, 40, 63, 65, 72, 74, 75, 76, 92, 94, 96, 110, 114, 116, 222, 224, 225	15A ,16A ,17A ,19A ,21A ,35A ,36A ,37A ,38A ,39A ,41A ,42A ,45A ,46A ,48A ,67A ,69A ,71A ,73A ,78A ,80A ,82A ,100A ,102A ,116A ,120A ,122A ,110D ,112D ,113D	12346	2964	148958	28
1sri-clean	11, 12, 13, 15, 17, 31, 33, 34, 35, 37, 38, 42, 62, 64, 71, 73, 74, 75, 77, 91, 93, 95, 109, 113, 115	23A ,24A ,25A ,27A ,29A ,43A ,45A ,46A ,47A ,49A ,50A ,54A ,56A ,75A ,77A ,79A ,86A ,88A ,89A ,90A ,92A ,106A ,108A ,110A ,124A ,128A ,130A ,118B ,120B ,121B	6827	2017	124187	25
1lke-clean	24, 27, 30, 31, 34, 35, 41, 43, 52, 54, 65, 82, 84, 89, 91, 95, 110, 116, 118, 120, 122	28A ,31A ,35A ,38A ,39A ,45A ,47A ,56A ,58A ,69A ,84A ,86A ,88A ,93A ,95A ,99A ,114A ,127A ,129A ,131A	15647	3260	135535	21
1lnm-clean	24, 27, 30, 31, 34, 35, 41, 43, 52, 54, 65, 80, 81, 82, 83, 84, 86, 91, 92, 95,110,119,121,123	28A ,31A ,35A ,38A ,39A ,45A ,47A ,56A ,58A ,60A ,69A ,73A ,84A ,86A ,88A ,90A ,93A ,95A ,99A ,114A ,127A ,129A ,131A	9306	2132	142313	24
1n0s-clean	28, 31, 45, 56, 58, 60, 69, 73, 84, 86, 88, 95, 97, 99, 100, 113, 114, 126, 127, 129, 131, 292	11A ,19A ,28A ,31A ,45A ,47A ,56A ,58A ,60A ,69A ,73A ,84A ,86A ,88A ,90A ,95A ,97A ,99A ,100A ,102A ,111A ,113A ,114A ,127A ,128A ,129A ,131A ,154A	11272	2171	126015	22
2XN3-clean	2, 5, 9, 218, 220, 222, 228, 230, 248, 251, 252, 254, 255, 256, 258	20A ,23A ,24A ,27A ,225A ,227A ,229A ,236A ,238A ,240A ,246A ,248A ,250A ,266A ,268A ,269A ,270A ,272A ,273A ,276A	1145	358	105834	15
2IZL-clean	11, 12, 13, 15, 31, 32, 33, 34, 35, 37, 38, 42, 65, 67, 74, 76, 78, 80, 96, 98,112,116,2 30, 231	23B ,24B ,25B ,27B ,29B ,43B ,45B ,46B ,47B ,49B ,50B ,54B ,56B ,75B ,77B ,79B ,86B ,88B ,90B ,92B ,106B ,108B ,110B ,124B ,128B ,130B ,118D ,120D ,121D ,124D	7597	1843	97588	24
4QAC-clean	95, 97, 150, 151, 152, 153, 154, 187, 189, 196, 198, 247, 264, 310, 325, 327	19A ,85A ,87A ,89A ,139A ,142A ,143A ,144A ,145A ,146A ,183A ,185A ,190A ,192A ,194A ,36B ,53B ,55B ,73B ,99B ,102B ,104B ,106B ,112B ,114B ,116B ,164B	4164	911	59939	16

2.8 Continued

Benchmark Complex	Designable Positions (Rosetta numbering)	Designable Positions (PDB numbering)	Total rotamers, accepted	Total rotamers, applied	Total rotamers, generated	Designable positions, count
4AFH-clean	98, 100, 102, 152, 153, 154, 155, 156, 192, 194, 201, 277, 325, 328, 341, 388	25A ,98A ,100A ,102A ,108A ,152A ,153A ,154A ,155A ,156A ,192A ,194A ,201A ,43B ,64B ,66B ,88B ,112B ,115B ,128B ,175B	3767	1227	92145	16
3CZ1-clean	32, 33, 37, 43, 46, 47, 48, 50, 51, 54, 56, 57, 99, 100, 102, 103, 106, 107, 114, 115, 116, 117, 122, 125	13A ,17A ,34A ,35A ,39A ,48A ,49A ,50A ,52A ,53A ,58A ,59A ,70A ,73A ,74A ,92A ,101A ,102A ,104A ,105A ,108A ,109A ,115A ,116A ,117A ,118A ,119A ,9B ,12B ,16B	2949	1257	111069	24
5t52-clean	43, 77, 78, 93, 94, 97, 98, 106, 122, 124, 125, 126, 127, 129	42A ,43A ,77A ,78A ,93A ,94A ,97A ,98A ,102A ,122A ,124A ,125A ,126A ,127A ,209A ,210A ,212A ,213A	5968	1013	63733	14
3oki-clean	21, 22, 27, 28, 31, 35, 36, 41, 44, 45, 48, 49, 52, 86, 87, 89, 90, 91, 92, 93, 94, 106, 109, 110, 113, 115, 120, 123, 126, 127, 201, 208, 212	264A ,267A ,268A ,273A ,274A ,277A ,281A ,287A ,290A ,291A ,293A ,294A ,295A ,298A ,299A ,302A ,332A ,333A ,335A ,336A ,337A ,338A ,339A ,340A ,344A ,352A ,355A ,356A ,359A ,361A ,366A ,369A ,372A ,373A ,390A ,447A ,454A ,458A ,473A	5090	1689	100259	33
4AFG-clean	99, 151, 152, 153, 154, 155, 191, 192, 193, 197, 198, 199, 200, 275, 277, 278, 326, 329, 337, 338, 339, 386	25A ,100A ,152A ,153A ,154A ,155A ,156A ,194A ,198A ,199A ,201A ,41B ,64B ,66B ,86B ,88B ,115B ,116B ,118B ,120B ,126B ,128B ,165B ,175B	1746	717	63564	22
4B5D-clean	101, 152, 153, 154, 155, 193, 200, 276, 296, 298, 326, 327, 328, 330, 336, 337, 338	98A ,100A ,102A ,152A ,153A ,154A ,155A ,156A ,194A ,198A ,199A ,201A ,203A ,41B ,64B ,66B ,68B ,86B ,88B ,115B ,116B ,117B ,118B ,120B ,126B ,127B ,128B ,165B ,174B ,175B	3738	761	62034	17
2QRY-clean	9, 10, 11, 65, 103, 107, 138, 144, 179, 180, 183, 196, 197, 199, 200, 203, 204, 225, 227, 261, 262, 263	27A ,29A ,30A ,59A ,83A ,84A ,120A ,121A ,125A ,128A ,156A ,161A ,162A ,197A ,198A ,201A ,205A ,213A ,214A ,215A ,217A ,218A ,221A ,222A ,243A ,245A ,275A ,279A ,280A ,281A	827	395	58642	22

Table 2.9: Forward Design Complexes Design Details. Statistics for complementary rotamers generated and applied to forward design of complexes generated with RosettaMatch for application ligands. Match scaffold PDB ID, scaffold description, complex designable positions in Rosetta numbering, ligand chemical component identified, and constraint file that yielded the match are provided. We also report the number of designable positions as well as the number of complementary rotamers generated, accepted (i.e. passed Rosetta energy and RMSD filters, see Methods), and applied to design for each benchmark binding site. Only the best 50 rotamers per residue type per position were applied top design.

Design	Scaffold (PDB ID: Chain)	Scaffold Description	Ligand	Binding Site Definition	DesignablePositions (Rosetta numbering)	Total rotamers, accepted	Total rotamers, applied	Total rotamers, generated	Designable positions, count
ATZ_1TP6	1TP6:A	1.5 A Crystal Structure of a NTF-2 Like Protein of Unknown Function PA1314 from Pseudomonas aeruginosa	ATZ	ATZ_0022-iter_0-fuzz_0-1_228_1014_2314	7, 8, 10, 12, 15, 16, 17, 30, 31, 34, 40, 42, 48, 53, 57, 64, 67, 69, 71, 88, 90, 92, 99, 101, 103, 105, 117, 120, 122	1414	437	73121	29
ATZ_1Q40	1Q40:A	Crystal structure of the C. albicans Mtr2-Mex67 M domain complex	ATZ	ATZ_0018-iter_0-fuzz_0-1_655_1670_2136	14, 15, 42, 48, 49, 50, 51, 67, 70, 77, 79, 103, 105, 133, 135, 136, 137, 138, 155, 156, 157, 158, 159	2715	727	103619	23
DOG_3ER7	3ER7:B	Crystal structure of NTF2-like protein of unknown function (YP_001812677.1) from EXIGUOBACTERIUM SP. 255-15 at 1.50 A resolution	DOG	DOG_0001-iter_0-fuzz_0-1_288_1136_2530	6, 7, 10, 11, 14, 23, 35, 54, 57, 61, 62, 63, 64, 67, 78, 79, 80, 81, 82, 92, 94, 95, 98, 111, 113, 116	10809	2592	165516	26
IBP_1JVX	1JVX:A	Maltodextrin-binding protein variant D207C/A301GS/P316C cross-linked in crystal	IBP	IBP_0007-iter_0-fuzz_0-1_759_1061_2710	11, 14, 15, 42, 43, 44, 45, 61, 62, 63, 65, 66, 111, 113, 153, 156, 210, 230, 258, 262, 331, 338, 341, 345	85	85	84936	24
IBP_1TUH	1TUH:A	Structure of Bal32a from a Soil-Derived Mobile Gene Cassette	IBP	IBP_0019-iter_0-fuzz_0-1_895_2639_2948	14, 17, 25, 35, 36, 37, 41, 43, 44, 47, 53, 56, 57, 59, 63, 68, 70, 87, 98, 100, 106, 117, 118, 119, 120, 126, 127, 130	1025	418	98950	28
IBP_3ECF	3ECF:A	Crystal structure of an ntf2-like protein (ava_4193) from anabaena variabilis atcc 29413 at 1.90 A resolution	IBP	IBP_0004-iter_0-fuzz_0-1_823_1135_2552	8, 9, 11, 13, 15, 16, 21, 23, 24, 26, 30, 41, 44, 47, 48, 50, 52, 55, 61, 77, 79, 91, 93, 100, 103	3753	1248	120051	25
IM4_1JKG	1JKG:A	Structural basis for the recognition of a nucleoporin FG-repeat by the NTF2-like domain of TAP-p15 mRNA nuclear export factor	IM4	IM4_0007-iter_0-fuzz_0-1_794_1164_1278	23, 24, 26, 28, 31, 34, 44, 46, 47, 51, 57, 60, 61, 62, 63, 64, 66, 67, 69, 71, 94, 100, 108, 110, 130, 134	3444	767	100199	26

2.9 Continued

Design	Scaffold (PDB ID: Chain)	Scaffold Description	Ligand	Binding Site Definition	DesignablePositions (Rosetta numbering)	Total rotamers, accepted	Total rotamers, applied	Total rotamers, generated	Designable positions, count
IM4_1Z1S	1Z1S:A	Crystal Structure of Putative Isomerase PA3332 from Pseudomonas aeruginosa	IM4	IM4_0017-iter_0-fuzz_0-1_742_1081_1151	17, 26, 29, 39, 58, 62, 65, 68, 69, 70, 71, 97, 104, 106, 108, 122, 126, 130, 131, 134	412	117	51352	20
IM4_3EMM	3EMM:A	X-ray structure of protein from Arabidopsis thaliana AT1G79260 with Bound Heme	IM4	IM4_0018-iter_0-fuzz_0-1_1192_1270_1360	25, 35, 51, 53, 63, 81, 83, 85, 87, 89, 91, 104, 106, 109, 112, 115, 118, 119, 120, 121, 131, 132, 134, 135, 137, 142, 145, 148	2302	737	88738	28
IM4_3GZB	3GZB:A	Crystal structure of putative SnoaL-like polyketide cyclase (YP_001182657.1) from Shewanella putrefaciens CN-32 at 1.44 A resolution	IM4	IM4_0002-iter_0-fuzz_0-1_1083_1296_1408	29, 30, 32, 34, 35, 36, 38, 41, 44, 53, 66, 67, 69, 70, 73, 74, 77, 79, 80, 82, 100, 102, 124, 126, 135, 140, 142	816	213	76121	27
LFN_2FNC	2FNC:A	Thermotoga maritima maltotriose binding protein bound with maltotriose.	LFN	LFN_0001-iter_0-fuzz_0-1_2065_2410_2917	7, 8, 9, 10, 34, 35, 39, 57, 59, 106, 148, 151, 154, 208, 225, 227, 229, 261, 296	2192	598	40354	19
LFN_2OWP	2OWP:A	Crystal structure of a cystatin-like fold protein (bxe_b1374) from burkholderia xenovorans lb400 at 2.00 A resolution	LFN	LFN_0001-iter_0-fuzz_0-1_1291_1909_2515	17, 20, 21, 23, 25, 32, 44, 47, 61, 67, 73, 76, 90, 92, 94, 103, 105, 107, 118, 121, 125	2086	547	49014	21
NPS_2RCD	2RCD:A	CRYSTAL STRUCTURE OF A PROTEIN WITH UNKNOWN FUNCTION FROM DUF3225 FAMILY (ECA3500) FROM PECTOBACTERIUM ATROSEPTICUM SCRI1043 AT 2.32 A RESOLUTION	NPS	NPS_0005-iter_0-fuzz_0-1_140_1081_2929	20, 21, 23, 24, 26, 28, 30, 32, 35, 38, 39, 47, 50, 52, 54, 60, 61, 63, 64, 65, 67, 69, 72, 74, 95, 106, 108, 110, 119, 122	6517	1869	130983	30
NPS_3GZR	3GZR:A	CRYSTAL STRUCTURE OF AN UNCHARACTERIZED PROTEIN WITH A CYSTATIN-LIKE FOLD (CC_2572) FROM CAULOBACTER VIBRIOIDES AT 1.40 A RESOLUTION	NPS	NPS_0005-iter_0-fuzz_0-1_47_259_1081	14, 15, 18, 21, 26, 36, 39, 40, 42, 44, 50, 53, 54, 56, 57, 58, 60, 61, 67, 88, 90, 95, 101, 106, 108, 126, 128	3974	1142	106224	27

2.6.2 Figures

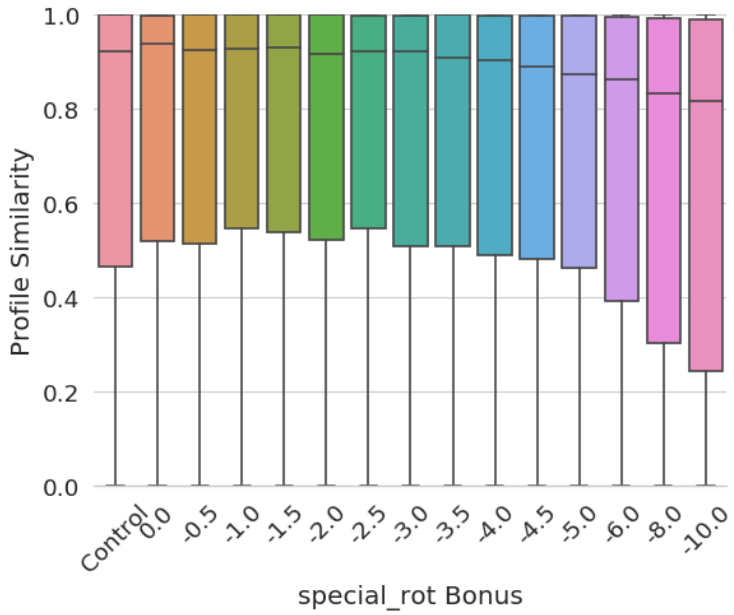


Figure 2.8: Profile similarity without >0.9 positions removed. Profile similarities as depicted in Figure 2.4 panel B, except profile similarities for all designable positions are included.

2.6.3 Files

S1 Appendix. Plots for all design metrics measured during forward design.

Plots for all design metrics measured during forward design. Design metric distributions measured for 5000 forward design trajectories of selected matches listed in Table 2.9. Existing Rosetta filters were applied to calculate binding strain (bindingstrain), interaction energy between the protein and ligand (residueie), packing statistics (Packstat), shape complementarity between the protein and ligand (shapecomplementarity), packing with RosettaHoles (holes), and the number of buried unsatisfied hydrogen bond donors/acceptors (heavyburiedunsats) for each design. Ligand solvent accessible surface area in $Å^2$ (ligand_sasa), number of hydrogen bonds made between the protein and ligand (hbond), and number of complementary rotamers incorporated during design (incorporated_specialrot) were also calculated.

S2 Appendix. Sequence Logos for binding site benchmark complexes at various special_rot bonuses.

Sequence logos generated for designable positions in existing complexes listed in S6 Table for various special_rot bonuses, where each logo represents 5000 design trajectories. A special_rot bonus of 0 indicates that complementary rotamers were added to the Packer, but received no score term bias. "Control" sequence logo was generated using the unmodified Packer. "Native" sequence logo shows the "correct" residue identity found in the native protein-ligand complex. Sequence logo positions are in Rosetta numbering.

S1 File. PDBs for binding site recovery benchmark.

PDB files for the protein-ligand complexes used for the binding site sequence recovery benchmark. Files have been cleaned and renumbered for compatibility with Rosetta, where position numbering corresponds to column "Designable_positions" in Table 2.8.

S2 File. Match constraints for existing complexes in match comparison.

RosettaMatch constraints (.cst) generated for existing complexes in the PDB for ligands in the Match comparison benchmark. Constraints consist of the three lowest-energy contacts in the binding site for the given complex in terms of the two-body Rosetta energies used to assemble contact pools (fa_rep, fa_atr, fa_elec, hbond_sc, fa_sol) and the hbond_bb_sc term to account for binding site backbone contacts with the ligand. Existing protein-ligand complexes were relaxed with the Rosetta FastRelax protocol with the full REF2015 energy function and starting coordinate constraints before determining constraint contacts.

S3 File. Top5000 RosettaMatch constraints for ligands in match comparison.

RosettaMatch constraints (.cst) generated for the 5000 lowest-energy composite binding site solutions found for application ligands listed in Table 2.4 and applied in the Match comparison benchmark.

S4 File. Composite binding site definition, RosettaMatch constraint file, and PDBs for matches carried forward through design.

PDB files for matches that were selected for forward design, the RosettaMatch constraint files (.cst) that yielded the match, and corresponding source composite binding site definition (.pdb) that was used to generate the RosettaMatch constraint file.

Chapter 3

Future Work

While our computational benchmarks demonstrate the viability of these methods in improving quality metrics pertinent to binding site design, validation ultimately requires design and testing binders in an experimental setting. We outline two approaches for validating designs generated using the methods in this work. The first strategy involves designing binders for the compound DFHBI as in Dou et al.[1] and screening for binding function in Escherichia coli with a fluorescent output. DFHBI is a GFP chromophore analog that is non-toxic, cell-permeable, and fluoresces upon binding when its internal degrees of freedom are constrained to a planar syn-conformation.[2] In addition, this compound is the chromophore for the Spinach aptamer that can be applied as a positive control. The second strategy involves designing chemically-induced dimerization systems for ligands of interest as our lab has previously demonstrated.[3]

Despite the importance of water in mediating interactions within ligand binding sites, the methods outlined in this work fail to recognize and incorporate waters as placing functional waters is a complex and ongoing research topic.[4,5,6] However, once these interactions are defined, it should be relatively straightforward to define Rosetta ResidueTypes that incorporate a residue-water interaction for application in composite binding sites and design with complementary rotamers.

Another interaction type not considered in this work was protein backbone contacts with the ligand. Backbone contacts were excluded from solutions for composite binding sites since an entire residue was included in the solution, and therefore sidechains of residues that contributed to a backbone-ligand interaction sterically occluded otherwise viable sidechain and backbone contacts with the ligand. This was still an issue with sidechain interactions with the ligand, but the

backbone is much smaller than the average sidechain considered for composite binding site solutions. Casting contact pool residues as virtual VariantTypes might provide a solution for this issue, where virtual atoms in Rosetta are not considered during scoring. Since we only care about the protein-ligand interactions and potential sidechain/backbone packing in composite binding sites, defining contact pool residues with virtual backbones/sidechains during simulated annealing could potentially resolve this issue.

It would be extremely interesting to apply composite binding sites in other productive contexts. In this work, we only generated composite binding sites composed of three residues. However, it is trivial to expand composite binding sites to an arbitrary number of residues. Composite binding sites of arbitrary size could serve as a starting point to design entire functional proteins from the inside-out. Adoption of a high-throughput matching algorithm[7] should enable screening of a greater number of more complex composite binding sites against larger scaffold libraries, ultimately permitting application of more stringent quality filters during this step. If only a subset of residues in a composite binding site are matched into a scaffold, loop modeling methods such as kinematic closure,[8] pull-into-place (an iterative modeling protocol that makes use of kinematic closure to precisely position functional residues on loops), or machine learning methods could be applied to build out a protein to accommodate the remaining composite binding site residues.

Finally, there are several improvements to software infrastructure that would expedite adoption and use of this method. There are several steps in the protocol, such as assembly of fragment contact pools, that would benefit from parallelization. At the moment, most processes are performed in series and do not benefit from additional available computing resources. Specifically, adapting this work to parallelization on cloud or highly-parallel computing resources would facilitate adoption by academic and industry researchers alike.

3.1 References

[1] Jiayi Dou et al. “De novo design of a fluorescence-activating β-barrel”. In: *Nature* 561.7724 (Sept. 2018), pp. 485–491. ISSN: 14764687. DOI: 10.1038/s41586-018-0509-0 (cit. on p. 55).

[2] Katherine Deigan Warner et al. “Structural basis for activity of highly efficient RNA mimics of green fluorescent protein”. In: *Nature Structural and Molecular Biology* 21.8 (2014), pp. 658–663. ISSN: 15459985. DOI: 10.1038/nsmb.2865 (cit. on p. 55).

[3] Anum A. Glasgow et al. “Computational design of a modular protein sense-response system”. In: *Science* 366.6468 (Nov. 2019), pp. 1024–1028. ISSN: 10959203. DOI: 10.1126/science.aax8780 (cit. on p. 55).

[4] Manuela Maurer and Chris Oostenbrink. “Water in protein hydration and ligand recognition”. In: *Journal of Molecular Recognition* 32.12 (2019), e2810 (cit. on p. 55).

[5] Alan P Graves et al. “A perspective on water site prediction methods for structure based drug design”. In: *Current topics in medicinal chemistry* 17.23 (2017), pp. 2599–2616 (cit. on p. 55).

[6] Gordon Lemmon and Jens Meiler. “Towards ligand docking including explicit interface water molecules”. In: *PloS one* 8.6 (2013) (cit. on p. 55).

[7] Tian Jiang et al. “An adaptive geometric search algorithm for macromolecular scaffold selection”. In: *Protein Engineering, Design and Selection* 31.9 (Nov. 2018), pp. 345–354. ISSN: 1741-0126. DOI: 10.1093/protein/gzy028 (cit. on p. 56).

[8] Daniel J Mandell, Evangelos A Coutsias, and Tanja Kortemme. “Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling”. In: *Nature methods* 6.8 (2009), pp. 551–552 (cit. on p. 56).

Chapter 4

Conclusion

The methods outlined in this work provide scientists with an expanded repertoire of computational tools to generate and design binding sites for arbitrary small molecules of interest. First, we remove the requirement that a complex for a target molecule must exist in the PDB to create a binding site definition: composite binding sites instead draw from protein-fragment interactions in the PDB for fragments that compose the target molecule to generate hundreds of thousands of composite binding sites de novo. Second, we augment Rosetta's design machinery with observed protein-fragment interactions and bias incorporation of these interactions during design to compensate for shortcomings in Rosetta's energy function when scoring protein-ligand interactions. Finally, we demonstrate that the combination of the methods introduced in this work result in designed binders with improved metrics that are indicative of design success. While there are several potential avenues for building upon this work, we have outlined several new strategies that can be immediately applied to design proteins that create novel protein-ligand interfaces and perform new functions. These methods can be applied beyond straightforward ligand binding and help inform design efforts for enzyme design and other protein sensing actuators.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
James Lucas
D98B18A0780E49F...
Author Signature

3/7/2020
Date